文章编号：2095-2163(2023)01-0171-07 中图分类号：TP391.4 文献标志码：A

基于 YOLOv3-SE-RE 模型的羊姿态识别

李小迪，王天一
（贵州大学 大数据与信息工程学院，贵阳 550000）

摘 要：为了实现养殖场环境下羊只个体的有效识别及对羊只姿态进行迅速判断，针对现有模型识别精度差，效率低等问题，基于自建数据集，提出基于 YOLOv3 的改进模型。在主干网络 darknet53 中，首先引入通道注意力模块压缩激励网络，增强网络对重要通道的关注度，提高网络检测精度；其次，将网络中的部分残差模块用循环特征移位聚合模块替代，以提高检测速度和检测精度；采用余弦退火动态学习率替代原有学习率，在训练过程中进行动态微调，使网络能轻松跳出局部最优解。实验结果表明：在羊只检测与姿态识别任务中，YOLOv3-SE-RE 算法的平均精度（mean Average Precision，*mAP*）比原始 YOLOv3 算法的 *mAP* 提升了 9.98%，同时检测速度也得到提升。

关键词：目标检测；压缩激励网络；循环特征移位聚合模块；羊姿态识别

Goat posture recognition based on YOLOv3-SE-RE model

LI Xiaodi，WANG Tianyi
（School of big data and Information Engineering，Guizhou University，Guiyang 550000，China）

【Abstract】 In order to realize the effective identification of sheep individuals and the rapid judgment of sheep posture in the farm environment，aiming at the problems of poor identification accuracy and low efficiency of the existing models，an improved model based on YOLOv3 is proposed based on the self-built data set. In the backbone network darknet53，the channel attention module is introduced to compress the excitation network to enhance the network´s attention to important channels and improve the network detection accuracy；Secondly，some residual modules in the network are replaced by cyclic feature shift aggregation module to improve the detection speed and accuracy. At the same time，cosine annealing dynamic learning rate is used to replace the original learning rate，and dynamic fine-tuning is carried out in the training process，so that the network can easily jump out of the local optimal solution. The experimental results show that in the task of sheep detection and attitude recognition，the *mAP* of YOLOv3-SE-RE algorithm is 9.98% higher than that of the original YOLOv3 algorithm，and the detection speed is also improved.

【Key words】 Target detection；squeeze-and-excitation network；recurrent feature-shift aggregator；sheep posture recognition

0 引 言

随着智能信息处理技术的快步发展，羊养殖业方式从传统的个体散养模式逐渐转变为规模化、智能化养殖。在传统养殖方式中多采用人工观测法[1]和无线射频[2]的方式，对羊只个体进行目标检测。人工观测法需要耗费大量的人力和时间，不仅检测效率低，且检测错误率高；无线射频的方式需要额外的设备与同步的识别方法，一定程度上提高了养殖场的运营成本，影响养殖场的经济效益。传统的姿态识别多使用回归出精确关节点坐标的方式或无线传感网络来对姿态进行判别，在运动灵活的个体识别上可扩展性较差，且识别成本较高。

近年来，基于深度学习的目标检测[3]和姿态识别的方法，已成为国内外研究热点，其典型算法为 YOLO 系列算法。YOLO 系列算法利用了回归的思想，能够在原图片中的位置上，回归出目标检测的边框和目标的类别。为了实时对羊只状况进行了解，本文以羊只养殖场监控视频为研究对象，使用 YOLOv3[4-6]网络。YOLOv3 主干网络采用残差结构，其目的是为了防止连续下采样导致的特征丢失。但是该方法仍然保持了传统的卷积操作，带来了巨

基金项目：贵州省科学技术基金（黔科合基础 ZK[2021]一般 304，[2020]1Y254）。

作者简介：李小迪（1996-），女，硕士研究生，主要研究方向：图像处理、计算机视觉；王天一（1989-），男，博士，副教授，主要研究方向：量子通信、图像处理、计算机视觉。

通讯作者：王天一 Email：tywang@ gzu.edu.cn

收稿日期：2022-06-08

大的 flpos 计算量,虽保证了特征提取的多尺度性,但是却一定程度的增加了模型推理的时间。在监控视频的羊只检测与识别的视觉任务中,羊只整体呈现白色,RGB 3 通道值均接近 255,其大部分本体并不包含足够的视觉信息。同时,目标本体一定程度上存在周围环境的遮挡,即使是人眼视觉都很难辨认。从深度学习层面可以认为这种监督信号比较稀疏,给检测和识别任务增加了难度。

在人脑对于羊只的感知中,当一张图片中存在羊只时,人眼看到后会下意识地将注意力转移到羊只上,而忽略周围环境。此外,人眼并不需要看全羊只整体,通过羊只犄角,腿部等局部特征便可辨认目标,还可以结合周围环境的全局线索去推理出羊只的具体位置。另一方面,对于骨干网络来说,占据整张图中绝大部分还是周围单调的环境信息,计算机对于不同通道之间的特征会等价处理。

受此启发,为了实现骨干网络对于多尺度特征提取的性能,去除单调的环境对目标的干扰,同时减少模型的计算量,本文选择循环特征移位聚合器(Recurrent Feature-Shift Aggregator,RESA)[7]来替代 YOLOv3 主干网络中的 Resblock[8]部分 ,且增加了通道注意力[9]模块压缩激励网络(Squeeze-and-Excitation Networks, SENet),使网络在训练过程中将更多的权重转移到重要的通道信息上。此外,本文采用余弦退火学习率[10-11],在整个训练过程中控制模型的收敛性。通过在训练初期设置较大的学习率来避免陷入局部最优解,并在训练过程中逐渐降低学习率,使得模型能够稳定的学习,进而收敛到全局最优解。

1 网络结构

1.1 YOLOv3

如图 1 所示,YOLOv3 由 darknet53、特征金字塔(Feature Pyramid Network,FPN)及 YOLO Head 3 部分构成。

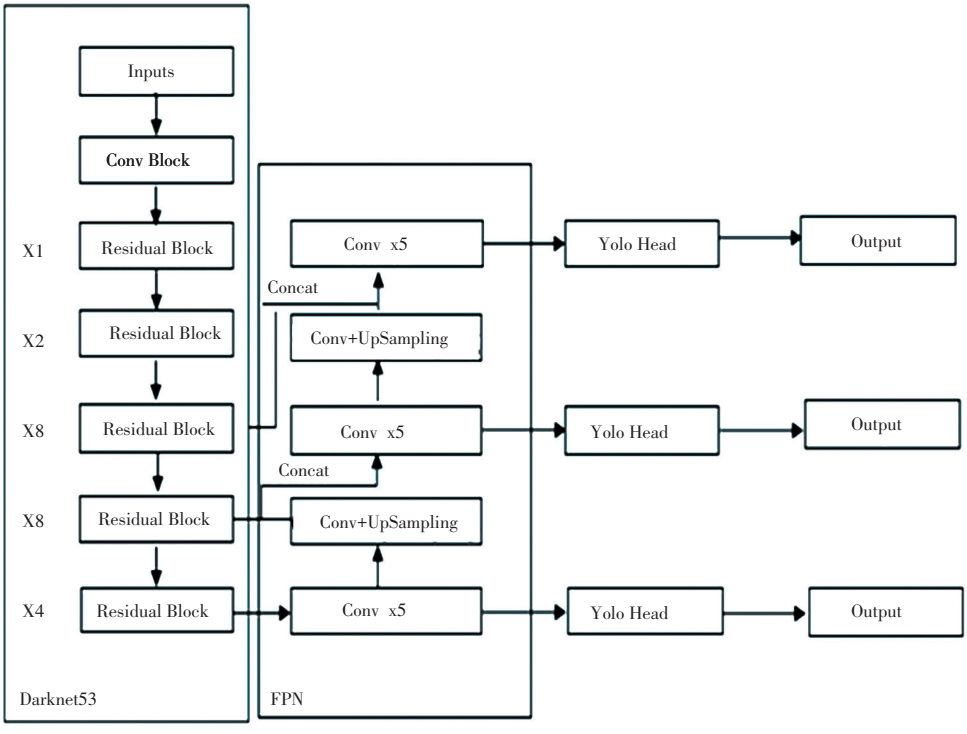


图 1 YOLOv3 网络结构

Fig. 1 Yolov3 network structure

Darknet53 被称作 YOLOv3 的主干特征提取网络,输入的图片首先会在 Darknet53 中进行特征提取。该模块由一个普通卷积模块和 5 个残差块组成,输入图片首先会被调整成 416×416×3 的大小,卷积过程对图片进行下采样处理,每经过一个卷积模块图片的宽和高就会被压缩至原图片的 1/2,通道数在卷积过程中不断扩张,以此获得一系列特征层,用来表示输入图片的特征。

FPN 被称作 YOLOv3 的加强特征提取网络,在主干部分获得的 3 个有效特征层,会在这一部分进行特征融合,特征融合的目的是结合不同尺度的特征信息。在获得 3 个有效特征层后,利用其进行

FPN 层的构建。

YOLO Head 实际上就是 YOLOv3 的分类器与回归器，其所做的工作就是进行分类预测与回归预测。因此，整个 YOLOv3 网络所作的工作就是特征提取-特征加强-获得预测结果 。

1.2　SENet

如图 2 所示为 SENet 的结构图，SENet 可以分为压缩（squeeze），激励 excitation）和重标定（reweight）3 个部分。

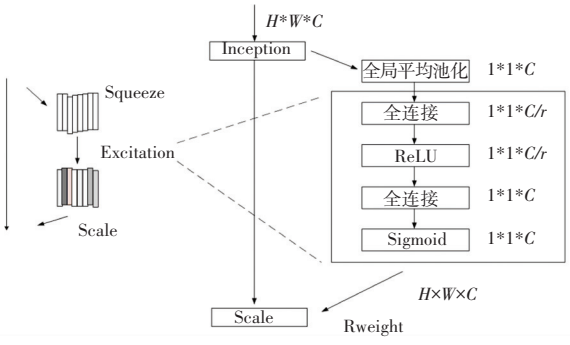


图 2　SENet 结构图

Fig. 2　Structure of SENet

输入一张大小为 $H*W*C$ 的特征图（H、W、C 分别为该特征图的高、宽、通道数），经过 squeeze 模块，将特征图顺着空间维度进行压缩，通过全局平均池化操作，将每个二维的特征通道变成一个实数，这个实数某种程度上具有全局的感受野，最终输出 $1*1*C$ 的特征图。输出的维度和输入的特征通道数相匹配，实现每个通道的所有特征求均值，旨在得到通道级的全局特征。

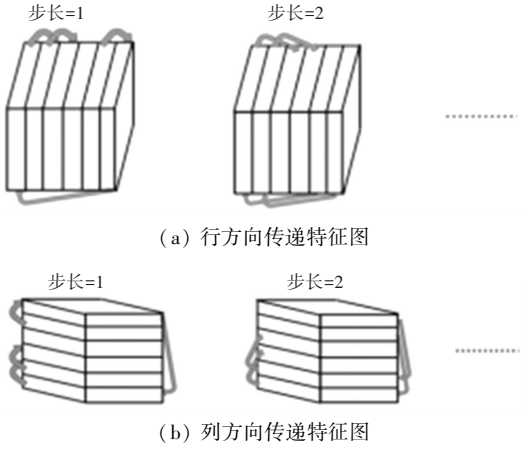
excitation 部分包括两个全连接层和两个激活函数。输入图片经过第一个全连接层时，通道数降为 C/r（r 为衰减因子，参照文献[12]中验证，r 取值为 16），然后使用 *ReLU* 激活函数激活；经过第二个全连接层时，恢复至 C 通道数，接着使用 *Sigmoid* 函数激活。基于通道间的相关性，每个通道生成一个权重，用来代表特征通道的重要程度。通过训练过程中学习权重，使得每一层通道获得非线性，即学习各个通道之间的主次关系。

最后，在 reweight 部分，将 excitation 输出的权重看做每个特征通道的重要性，通过乘法逐通道加权到之前的特征上，完成通道维度上对原始特征的重标定，从而实现不同通道特征重要性的区分。

1.3　RESA 算法

RESA 算法是通过一个介于编码器（用于特征提取）和解码器（用于目标恢复）之间的 RESA 模块，将骨干网络中提取到的空间信息（局部信息与全局信息）进行聚合，使得原始特征得到增强。该模块在特征传递之前会将特征图中的特征进行切片，若要将特征进行左右传递，则先将特征图在列方向分为很多个切片，随后不同特征加权叠加。同理，在行方向上进行特征切片以及加权，左右及上下方向的信息传递均能够增强羊只不同特征部位以及周围环境的关联性。通过信息传递，理论上能够有助于目标的推理。从网络结构来看，RESA 模块采用了大量切片特征加权，因此这种非常规卷积的特征传递方式，能够减少时间消耗，且经过不同步长的特征迭代，最终输出的特征图上，每个像素整合了全局的每一处特征信息，能够有效防止传播过程中信息的丢失。

RESA 模块如图 3 所示，其中包含了 4 种子模块。分别为左->右、上->下、右->左、下->上，每部分模块均为 n 次迭代。图 3（a）为左->右模块 1 - n 次迭代的结构图，其中包含了不同步长下的信息传递示意图。特征图被纵向分为许多切片，当步长为 1 时，由左数第一个切片的特征经卷积操作后，叠加至第二个切片。同理，当步长为 2 的时候会叠加至第三个切片，以此类推。图 3（b）为从下->上的模块，处理过程同上。



（a）行方向传递特征图

（b）列方向传递特征图

图 3　RESA 模块结构图

Fig. 3　Structure diagram of RESA module

1.4　YOLOv3-SE-RE 模型

本文在 YOLOv3 的主干网络上进行了模型的优化，当第一次完成特征层提取后，嵌入 SENet 模块，如图 4 所示。首先，输入图片的尺寸会被调整为 416×416×3，经过一次 1×1 卷积后，特征图的通道数得到扩展，尺寸变为 416×416×32；随后，特征图进入 SE 模块进行压缩，经过全局平均池化操作后，特征图的大小被压缩为 1×1×32；经过全连接层，特征图的大小变为 1×1×2，衰减因子 r 为 16。使用 *ReLU* 激

活函数进行激活,此时的通道数不变;再经过一层全连接层,恢复通道数为 32;最后使用 *Sigmoid* 函数进行激活,此时每个通道都分配到了不同的权重;再通过乘法,逐通道加权到之前的特征上,权重值越大,说明网络对该通道的关注度越高;其次,去除 YOLOv3 主干网络中的残差结构,将分配好权重的特征图输入到 RESA 模块中,进行行方向和列方向的切片处理,使得空间通道特征得以丰富。

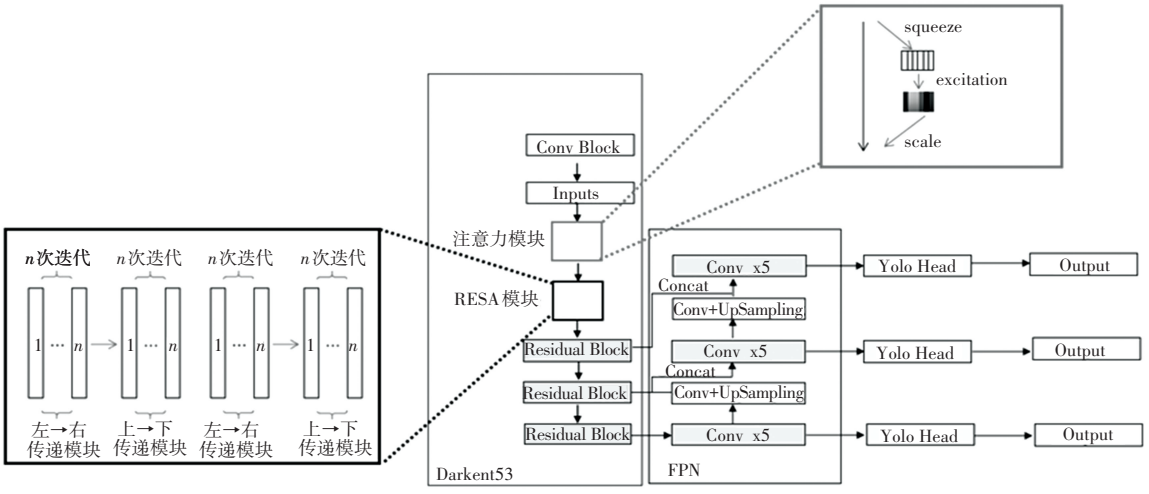


图 4 YOLOv3-SE-RE 模型结构图

Fig. 4 Structure diagram of Yolov3-SE-RE model

2 余弦退火学习率

在训练网络时,学习率会随着训练而发生变化。在训练网络的后期,若学习率过高,则会造成损失的振荡,而学习率衰减过快,则会造成模型收敛变慢的情况。由于模型在训练初期对于图像是完全未知的,即模型对于像素信息的理解相当于均匀分布,因此训练初期模型非常容易陷入过拟合。基于此,本文采用余弦退火方式对学习率进行调整,余弦退火学习率整体符合余弦函数的变化方式。余弦函数中,随着 x 的变化,函数值先缓慢下降然后加速下降,以此为一个周期循环。当模型经过几个轮次的训练后,逐渐对于数据集有所了解,此时需要降低学习率,使得模型能够稳定的学习,从而向着全局最优解去收敛。这种下降模式与学习率结合,能轻松让模型跳出局部最优解。学习率定义如下:

$$\eta_t = \eta_{\min}^i + \frac{1}{2}(\eta_{\max}^i - \eta_{\min}^i)(1 + \cos(\frac{T_{cur}}{T_i}\pi))$$

其中,i 表示索引值,即运行轮次,$\eta_{\max}^i$、$\eta_{\min}^i$ 表示学习率的最大、最小值。

这两个值限制了学习率的范围,使学习率能够在一定范围内衰减。T_{cur} 表示当前执行了多少个轮次(epoch),由于 T_{cur} 在每个批次(batch)运行后将会更新,而此时的 epoch 还没有执行完,因此 T_{cur} 可以为小数。T_i 表示第 i 次运行时总的 epoch 数。本文中,模型的初始学习率设置为 0.01,随着 epoch 的增加,学习率按照余弦规律减小,开始下降速度缓慢,当训练到第 20 个 epoch 时,学习率下降速度变快,最终大小为 5×10^{-4}。

3 实验

3.1 数据集

3.1.1 数据集构成

为了获取高质量的羊只图片,需要对获取到的羊舍监控视频进行预处理操作,首先将获得的3 047 个监控视频进行手动裁剪,裁剪出合适的角度后手动删除无效片段。由于羊只在监控视频中多出现站立和坐卧的姿势,于是将裁剪出的有效监控按照羊的姿势分为站立(stand)和坐卧(lie down)两个类别,共计 309 个有效监控视频。然后对有效监控视频进行关键帧的提取,由于羊只在羊舍内的活动范围较小,在短时间内羊只的姿态不会发生明显的变化,因此每隔 50 帧提取一张关键帧图片。此外,通过手动删除关键帧中羊只肢体不全、遮挡严重、无羊只等无效图像,最终得到 808 张羊只站立图片和 1 192张羊只坐卧图片,共计 2 000 张图片。

3.1.2 数据集的标注

本文使用 Labelimg 标注工具对羊只数据集图像进行标注,标注过程中为了不引入太多的背景,只对羊只主体进行标注,数据集按照训练集和测试集

8∶2 的方式划分，模型优化前后均在相同的数据集下训练测试 40 轮次。

3.2　实验环境

实验均在 Ubuntu18.04.4LTS 操作系统上进行，python3.7，tensorflow2.0 深度学习框架，cpu 为 i7-9700，显卡为 RTX 2080Ti，使用 Labelimg 对自建数据集进行标记，总共 2 000 张图片，按照 8∶2 的比例划分为训练集和测试集。

3.3　实验结果分析

本文将基于通道注意力机制的 YOLOv3 模型称为 YOLOv3-SE 模型、基于 RESA 算法的 YOLOv3 模型称为 YOLOv3-RE 模型、基于通道注意力机制及 RESA 算法的模型称为 YOLOv3-SE-RE 模型。

3.3.1　*mAP* 及检测速度对比

将上述 4 个模型在自建数据集上训练 40 个 epoch 后，对训练好的模型进行测试实验。测试集图片共 400 张，在相同的测试集下分别测试了算法优化前后模型的推理速度，即前向传播一次，推理一张图片所用的时间及模型在测试集上的 *mAP* 值见表 1。

表 1　YOLOv3 算法与改进后系列算法在测试集上的实验结果

Tab. 1　Experimental results of YOLOv3 algorithm and a series of improved algorithms on the test set

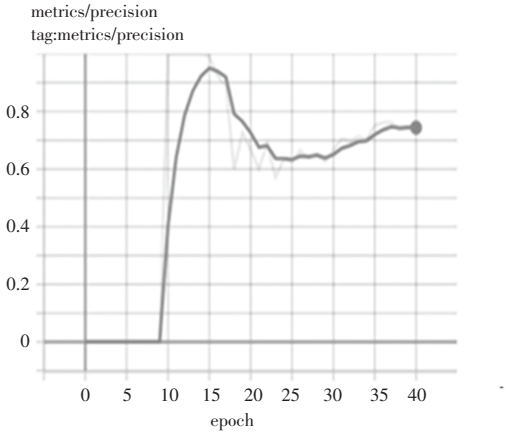
模型名称	测试集数量 /张	*mAP* /%	单张推理速度/fps
YOLOv3	400	86.02	29
YOLOv3-SE	400	88.37	30
YOLOv3-RE	400	91.26	32
YOLOv3-SE-RE	400	96.00	34

由此可见，YOLOv3 模型的 *mAP* 为 86.02%，YOLOv3-SE 模型的 *mAP* 达到了 88.37%，相比 YOLOv3 模型增加了 2.35%；YOLOv3-RE 模型的 *mAP* 达到了 91.26%，相比于 YOLOv3 模型增加了5.24%。

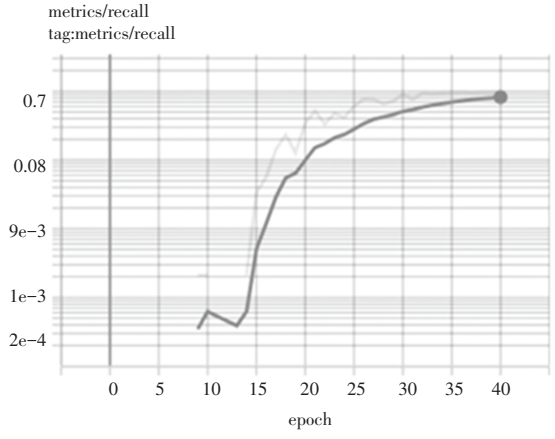
实验表明，增加了注意力机制的 YOLOv3 模型，以及增加了 RESA 模块的 YOLOv3 模型，目标检测均值平均精度略高于 YOLOv3 模型。而结合了两个模块的 YOLOv3-SE-RE 模型的目标检测的均值平均精度为 96%，明显高于 YOLOv3 模型，且单张图片推理速度也明显高于 YOLOv3 模型。

3.3.2　精确度与召回率对比

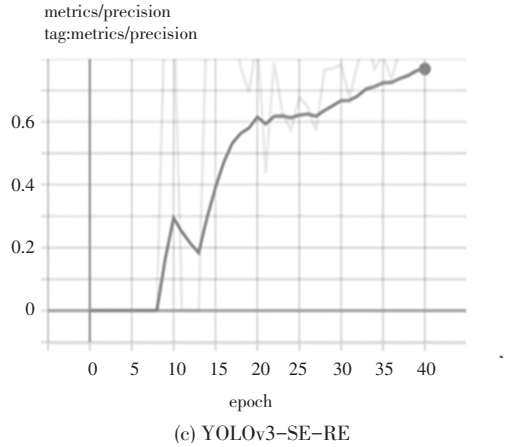
模型优化前后，精确度与召回率的对比结果如图 5 所示。



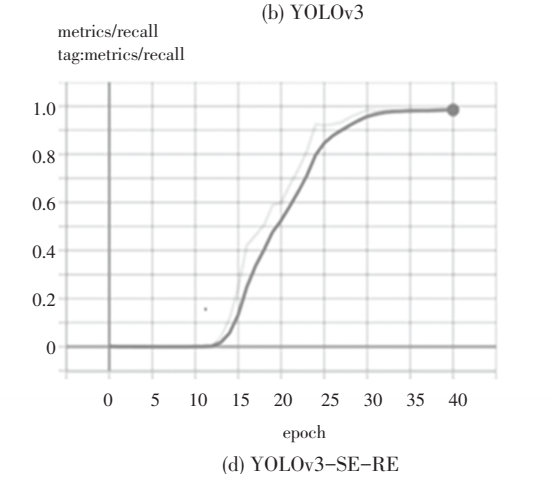
(a) YOLOv3



(b) YOLOv3



(c) YOLOv3-SE-RE



(d) YOLOv3-SE-RE

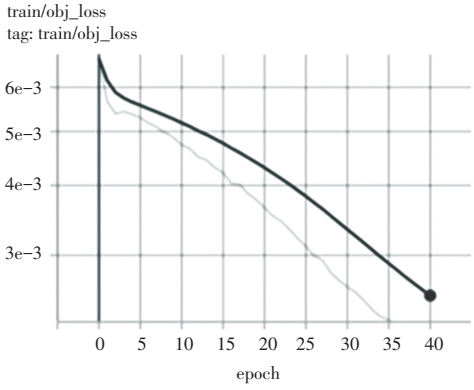
图 5　精确度与召回率比较

Fig. 5　Comparison between accuracy and recall

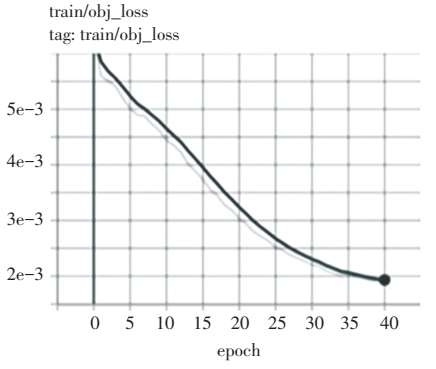
由图 5(a)、(b)所见,YOLOv3 模型的精确率与召回率稳定在 0.75 与 0.7,当模型在训练到第 15 个 epoch 时,精确度达到最高,然后开始下降。而 YOLOv3-SE-RE 模型两项测试指标在模型训练至 40 epoch分别提升至 0.79 与 0.97。由此可见,无论是精确率还是召回率,优化后的模型更加平稳,且收敛速度更快。

3.3.3 Loss 的对比

模型在训练过程中损失函数可视化结果如图 6 所示。由图 6(a)可见,obj_loss 在第 40 个 epoch 时损失达到了 2.5e-3。而从模型优化后的损失函数曲线(图 6(b))可以看到,当模型训练到第 40 个 epoch 时,损失可降低至 2e-3。由于余弦退火学习率的加入,使得整个训练过程中损失函数收敛的更快且更加平滑。



(a) YOLOv3



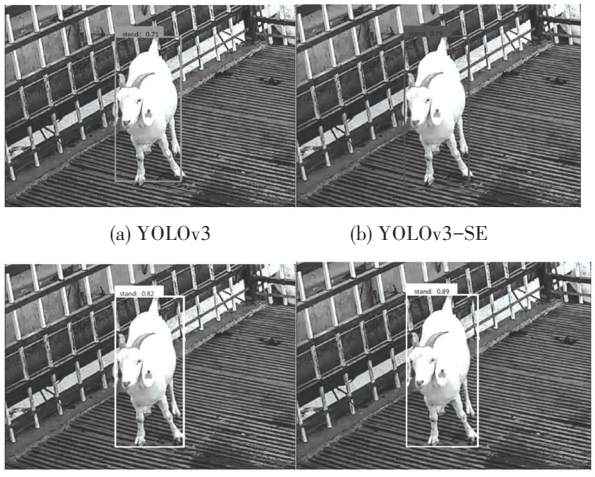
(b) YOLOv3-SE-RE

图 6 损失函数比较

Fig. 6 Loss function comparison

3.3.4 站立姿态检测效果对比

如图 7 所示,在站立姿态下,与 YOLOv3 算法相比,YOLOv3-SE 算法的预测框将右下角羊蹄完整框入其中,且预测框紧贴羊只个体;YOLOv3-RE 算法预测框将羊只的左右两只羊蹄完整框入其中,且预测框紧贴羊只个体;YOLOv3-SE-RE 算法预测框将羊只个体完整框入其中,且预测框紧贴羊只个体且没有将过多的背景框入其中。实验证明,通道注意力机制,可以使模型更加关注包含重要信息的通道,减少对背景的关注度。RESA 算法可以增加羊只不同部位之间的关联性,通道注意力机制与 RESA 模块均可以提升模型的检测精度。



(a) YOLOv3 (b) YOLOv3-SE

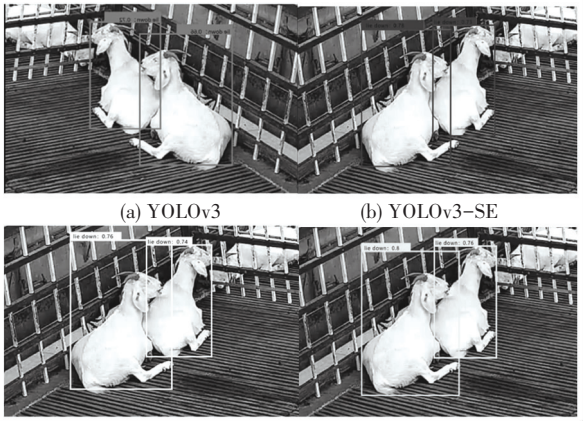
(c) YOLOv3-RE (d) YOLOv3-SE-RE

图 7 站立姿态检测效果图

Fig. 7 Effect diagram of standing posture detection

3.3.5 坐卧姿态检测效果对比

如图 8 所示,在坐卧姿态中,与 YOLOv3 算法相比,YOLOv3-SE 算法的预测框更完全的将羊只框入其中,且预测框紧贴羊只个体没有框入过多背景;YOLOv3-RE 算法预测框的准确率以及置信度均略高于 YOLOv3 模型;YOLOv3-SE-RE 算法预测框,将两只紧贴的羊只个体完全框入其中,预测框紧贴羊只个体并没有将过多背景,且置信度也有所提高。实验证明,在坐卧姿态下,改进后的网络识别效果明显优于原模型。



(a) YOLOv3 (b) YOLOv3-SE

(c) YOLOv3-RE (d) YOLOv3-SE-RE

图 8 坐卧姿态检测效果

Fig. 8 Sitting and lying posture detection effect

4　结束语

本文通过对 YOLOv3 算法的优化，实现了对视频监控中羊只姿态的高效识别。首先在主干网络 darknet53 中增加通道注意力机制，增加不同通道的特征相关性，让网络重点关注权重值较大的通道信息，以提高网络的检测精度。其次通过增加 RESA 模块，对特征图进行行方向和列方向的切片和聚合，增加目标检测物体不同部位之间的关联性，同时提高了检测精度和速度。实验结果表明，YOLOv3-SE-RE 模型在检测精度和检测速度上都超过了原始 YOLOv3 模型，对于不同姿态的识别，效果也有明显的优化，本应用在智能养殖方面有较好的应用前景。

参考文献

[1] NEETHIRAJAN S, TUTEJA S K, HUANG S T, et al. Recent advancement in biosensors technology for animal and livestock health management[J]. Biosensors and Bioelectronics, 2017, 98: 398-407.

[2] ROBERTS C M. Radio frequency identification (RFID)[J]. Computers & security, 2006, 25(1): 18-26.

[3] 许德刚，王露，李凡. 深度学习的典型目标检测算法研究综述. 计算机工程与应用，2021，57(8)：10-25.

[4] REDMON J, FARHADI A. Yolov3: An incremental improvement [J]. arXiv preprint arXiv:1804.02767, 2018.

[5] BENJDIRA B, KHURSHEED T, KOUBAA A, et al. Car detection using unmanned aerial vehicles: Comparison between faster r-cnn and yolov3[C]//2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS). IEEE, 2019: 1-6.

[6] 刘生智，李春蓉，刘同金，等. 基于 YOLOV3 模型的奶牛目标检测[J]. 塔里木大学学报，2019，31(2)：85-90.

[7] ZHENG T, FANG H, ZHANG Y, et al. Resa: Recurrent feature-shift aggregator for lane detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(4): 3547-3554.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[10] LOSHCHILOV I, HUTTER F. Sgdr: Stochastic gradient descent with warm restarts[J]. arXiv preprint arXiv:1608.03983, 2016.

[11] HE T, ZHANG Z, ZHANG H, et al. Bag of tricks for image classification with convolutional neural networks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 558-567.

[12] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

(上接第 170 页)

[6] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [EB/OL]. 2018, arXiv:1802.05365.

[7] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[8] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding [J]. Advances in neural information processing systems, 2019, 32.

[9] SONG K, TAN X, QIN T, et al. Mpnet: Masked and permuted pre-training for language understanding[J]. Advances in Neural Information Processing Systems, 2020, 33: 16857-16867.

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

[11] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[12] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]// Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001: 282-289.

[13] ZHANG H, ZONG Y, CHANG B, et al. 面向医学文本处理的医学实体标注规范 (medical entity annotation standard for medical text processing)[C]//Proceedings of the 19th Chinese national conference on computational linguistics. 2020: 561-571.

[14] LUO L, YANG Z, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition [J]. Bioinformatics (Oxford, England), 2018, 34(8): 1381-1388.

[15] YAN R, JIANG X, DANG D. Named Entity Recognition by Using XLNet-BiLSTM-CRF [J]. Neural Processing Letters, 2021, 53(5): 3339-3356.

[16] 许力. 基于 BERT 和 BiLSTM-CRF 的生物医学命名实体识别[J]. 计算机工程与科学，2021，43(10)：1873-1879.