Software Engineering and Applications 软件工程与应用, 2023, 12(1), 113-123
Published Online February 2023 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2023.121012

Hans汉斯

基于YOLO特征检测模型的非法溜宠物识别

黄学雷，翟昕宇

上海理工大学，上海

收稿日期：2023年1月21日；录用日期：2023年2月17日；发布日期：2023年2月24日

摘　要

非法溜宠物现象的识别属于细粒度图像分类的一种，针对复杂的生活环境，使用传统图像分类方法或者单纯的卷积神经网络来进行非法溜宠物现象的识别，会出现准确率偏低的情况。本文基于深度学习中实例分割方法来实现对溜宠物行为的识别，通过检测到的图像中目标的信息、目标与目标关系信息来实现对是否是非法溜宠物的现象判断。该方法是对2020年Glenn Jocher发表的实例分割模型YOLO (you only look once)的基础上进行改进，主要针对主干特征提取网络改为对细粒度图像更为友好的SENet特征网络。并将最深的stage部分的特征网络改为SPPCSPC模块用于优化特征提取精度。针对多种犬类和大量不同的现实环境的搭配来对整个网络进行训练。通过实际的公园活动场景中的识别表明，改进后的网络精度上变化不大，且很好地满足了实际生活中对于这种溜宠物现象的准确识别需要。

关键词

溜宠物识别，深度学习，YOLO

Recognition of Illegal Pet Walking Based on YOLO Feature Detection Model

Xuelei Huang, Xinyu Zhai

University of Shanghai for Science and Technology, Shanghai

Received: Jan. 21st, 2023; accepted: Feb. 17th, 2023; published: Feb. 24th, 2023

Abstract

The recognition of illegal pet walking is a kind of fine-grained image classification. For complex living environments, using traditional image classification methods or simple convolutional neural networks to recognize illegal pet walking will lead to low accuracy. This paper realizes the recognition of pet walking behavior based on the case segmentation method in depth learning, and judges whether it is illegal pet walking phenomenon by detecting the information of the target in

文章引用: 黄学雷, 翟昕宇. 基于 YOLO 特征检测模型的非法溜宠物识别[J]. 软件工程与应用, 2023, 12(1): 113-123.
DOI: 10.12677/sea.2023.121012

the image and the relationship between the target and the target. This method is an improvement on the case segmentation model YOLO (you only look once) published by Glenn Jocher in 2020. It is mainly aimed at changing the backbone feature extraction network to the SENet feature network that is friendlier to fine grained images. The feature network of the deepest stage is changed into SPPCSPC module to optimize the feature extraction accuracy and train the whole network according to the combination of a variety of dogs and a large number of different real environments. The identification in the actual park activity scene shows that the accuracy of the improved network has little change, and meets the needs of accurate identification of this pet walking phenomenon in real life.

Keywords

Pet Walking Recognition, Deep Learning, YOLO

1. 引言

我国拥有极为庞大的养宠物的群体，每日的溜宠物已经成为大多数有宠物家庭的生活中必不可少的部分。它极大地丰富了人们的日常生活。然而目前来说，对于用户溜宠物中一些违规现象的识别只能通过人为的识别或者违规现象发生后通过查看相关录像来实现对于违规现象的惩处。溜宠物现象的人为识别的人力成本以及违规现象惩处的滞后性对于溜宠物现象中违规行为识别都有很大的弊端。

2021 年 1 月 22 日，十三届中国人大常委会表决通过了新修订的《中华人民共和国动物防疫法》。溜宠物的违规行为纳入了法律的监管。人为识别以及惩处的滞后性对于溜宠物违规现象的监管就显得有些力不从心。因而自动、准确地识别溜宠物中的违规行为的研究更能符合时代的需要以及监管的需求了。

随着机器视觉的发展，对于动物的识别检测也迈向了一个新的台阶。从最初的将所有情况输入系统中以实现目标检测的算法系统，到之后人工设计的经典机器算法，例如 Sift (Scale-invariant feature transform)、HOG (Histogram of oriented gradient)、Harris 等算法[1]对目标进行特征提取，然后输入到 SVM (Support Vector Machine)等分类器中进行目标的检测，再到 2012 年 CNN 的崛起，实现了深度学习和计算机视觉相结合的阶段，迎来了目标检测算法的更深层次的研究。如今通过深度学习网络实现的特征提取能够实现更高的检测精度和特征提取能力，从 2014 年 rgb 提出的 R-CNN 检测算法，开始了双步(Two-stage)检测算法的研究，这个时期代表作有 Fast-Rcnn、Faster-Rcnn 等[2]，先通过 RPN 网络获得先验框，然后通过建议框对提取到的特征层进行截取。最后通过回归预测(box-regressing)对建议框进行调整，从而实现对目标的检测。2014 年 YOLO 的出现将双阶段(two-stage)的检测算法融合成了单阶段(one-stage)的检测算法。这一时期的代表作便是 YOLO 系列，如 YOLOv1，YOLOv2，YOLOv3 等[3]。因为单阶段的检测算法无需提取建议框而是直接输入到网络中生成目标的类别和位置，因此速度上比双阶段的检测算法更快。

上述的研究主要还是停留在大型物体识别研究上。而对于现实生活中溜宠物行为的识别更加复杂，除了准确识别目标对象之外，还需要准确识别宠物与人的关系，并综合所有复杂信息才能准确判断是否是违法的溜宠物。

本文提出了一种基于 YOLO 的改进的宠物违规现象的识别算法。使用界内已经十分成熟的大规模数据集 VOC_2017 上训练改进的 SENet 特征提取网络进行特征提取。然后对复杂的 59 种现实环境针对性的手动收集和制作部分数据集，再输入到网络中进行针对性的训练以达到精准检测的目的。

2. 相关理论

2.1. Backbone 介绍

2.1.1. Anchor

Anchor Box [4]是预先设定的一系列边框，在进行网络训练时，以真实的边框位置相对于预设边框的偏移来构建。预设定的边框可能的位置“框”出目标来，然后再在这些预设定的边框的基础上进行调整。

我们将输入的图片分割为 38 * 38 块区域，每一个区域由 9 个先验框负责进行识别。先验框会对这 38 * 38 个区域进行匹配，也就是一旦认为这些区域里存在目标，就让左上角点所携带的先验框去负责检测。

对于任何一个真实框，本文的算法不再使用以往的 IOU 正样本[5]匹配的方式进行选取。而是直接采用高宽比进行匹配。即使用真实的框和图 1 中 9 个不同大小的先验框计算高宽比。如果真实框与先验框的高宽比例大于设定的阈值，则说明该真实框和该先验框匹配度不够，将该先验框认定为负样本。

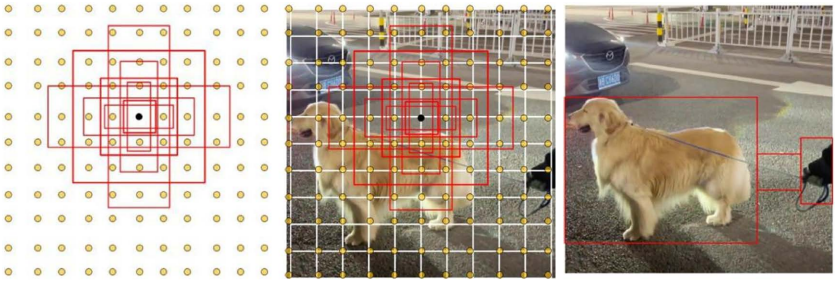

Figure 1. Anchor box
图 1. Anchor box

2.1.2. 下采样过渡模块

在常见的卷积神经网络中，常见的下采样过渡模块(如图 2)是一个卷积核大小为 3 * 3、步长为 2 * 2 的卷积或者步长为 2 * 2 的最大池化。通过卷积进行下采样的信息融合较好，而采用池化进行下采样则可以一定程度上忽略目标的倾斜、旋转等位置变化。从而避免网络的过拟合，提高了鲁棒性。

综合上述两种下采样方式的优点，组成了本文网络中的下采样模块，通过分别池化和卷积的下采样并将其特征进行堆叠。

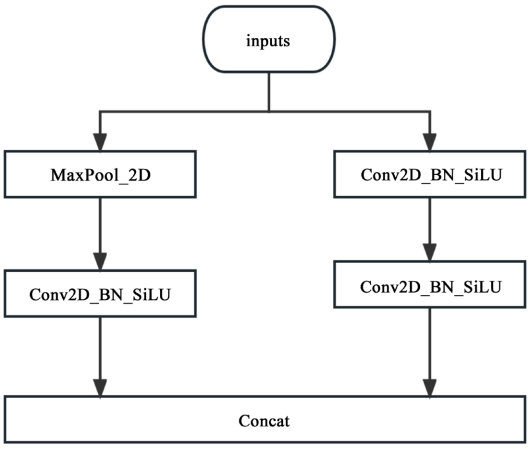

Figure 2. Down sampling transaction module
图 2. 下采样过渡模块

2.2. FPN 特征金字塔

在卷积网络中，随着网络深度的增加，特征图的尺寸也越来越小，语义信息也越来越抽象。浅层的语义信息较少，但是目标位置相对准确，目标位置比较粗略，导致小物体容易检测不到。FPN14 则是自顶向下地处理特征图并通过横向连接的方式融合底层的具有较少语义信息的特征图和高层的具有丰富语义信息的特征图，同时没有牺牲表达能力、速度和资源的消耗。

通过融合浅层到深层的特征图(如图 3)，从而充分利用各个层次的特征。

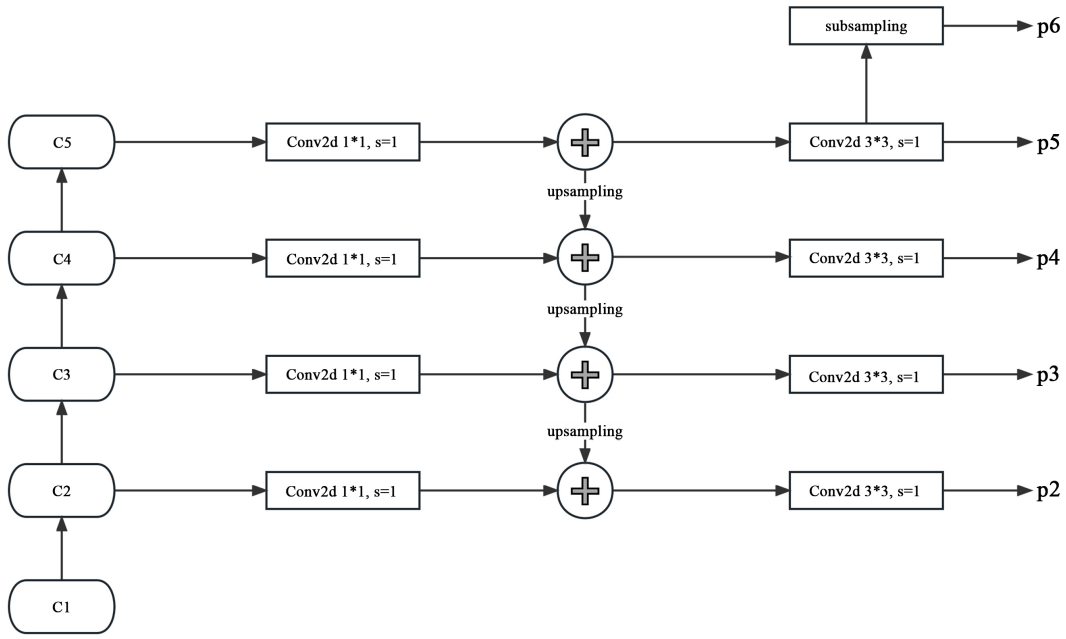

Figure 3. FPN feature pyramid
图 3. FPN 特征金字塔

如图 3 所示，在得到 c2、c3、c4、c5 这些特征图之后，首先将 c5 进行 1 * 1 卷积将其通道值变为 256，然后进行 2 倍上采样，此时得到的特征图的高和宽与 c4 是一样的，但是注意，c4 的 channel 值和上采样得到的特征图的 channel 值不太一样，所以 c4 也会先进性依次 1 * 1 的卷积将 channel 变成 256，然后和该特征图依次堆叠，最终得到融合的特征图。

总的来说就是，自顶向下，c5、c4、c3、c2 依次进行横向的 1 * 1 卷积，再与上一层上采样的结果进行矩阵加法操作的结果。

2.3. SPPCSPC 模块

SPP 模块(见图 4)的作用是能够增大感受野，使得算法适应不同的分辨率图像，它是通过最大池化来获得不同的感受野的。

我们可以通过再第一条分支中，经过 maxpool 的四条分支。分别是 5、9、13、1。这四个不同的 maxpool 就代表了他能够处理不同的对象，也就是说，这四种不同的尺度的池化有四种感受野，主要是区别大目标和小目标。

CSP 模块，首先将特征分为两部分，其中的一个部分进行常规的处理，另外一个部分进行 SPP 结构的处理，最后把这两部分进行合并，如此可以减少近一半的计算量，使得速度变快且精度增加。

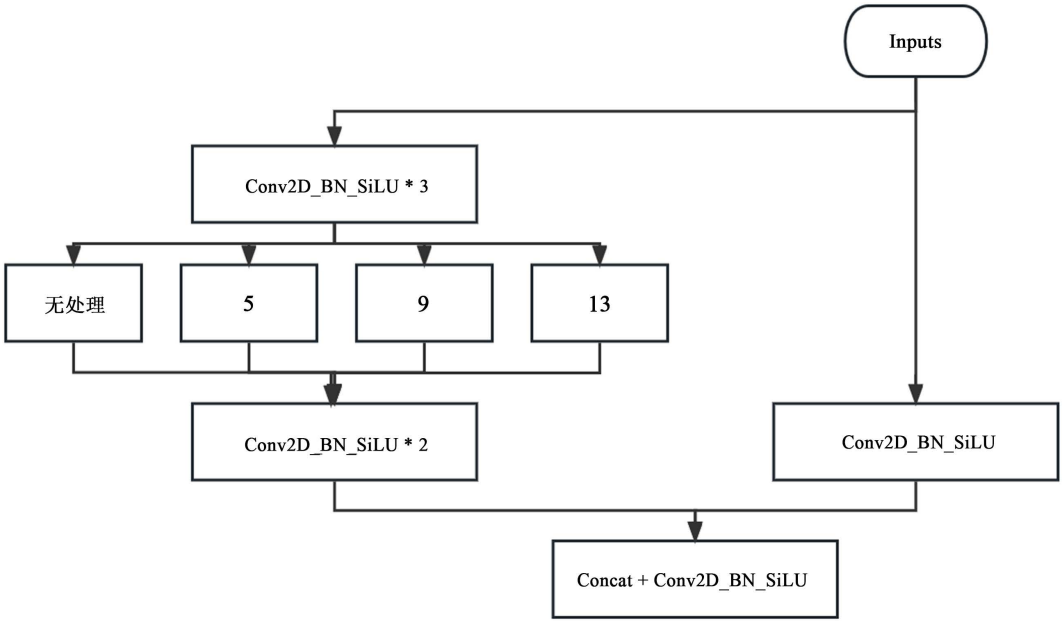

Figure 4. SPPCSPC module
图 4. SPPCSPC 模块

3. 违规溜宠物现象识别

3.1. 基本思想

基于合法溜宠物的标准，我们需要对宠物品种，是否拴绳，溜宠物过程中宠物排泄是否清理三个方面进行准确识别。

对于宠物分类时，由于不同的宠物，外观具有一定的相似性，同一种宠物毛发颜色、姿势不同等原因，因此需要充分包含目标纹理中的细节信息，为了能在得到全局视觉特征信息的同时，还能够得到图像关键区域的细节特征共同进行对比分析，本文提出了如图的实验过程。

1) 将特征提取网络在大型公共数据集 VOC2017 上训练获取相关的预训练的神经网络
2) 加入 SPPCSPC 模块
3) 更改融合点的网络结构，让整个输入的尺寸和通道数满足第二阶段的特征金字塔的需要
4) 将整个数据集按照 7:2:1 的比例分成训练集，验证集和测试集
5) 将训练后得到的 YOLOHead 进行解码处理
6) 使用得分抑制选取置信度最高的框
7) 使用非极大抑制算法以防同一个种类的预测框的堆积
8) 将测试集送入网络，计算准确率

最后通过 YOLO 预测框限制检测区域增加预测精度，并对人、绳子、宠物、粪便等四种目标的关系最终按照权重比例计算是否是违法溜宠物现象。

3.2. SENet

我们将原本 YOLO 中用于提取特征网络主要部分换成了 2017 年由 Momenta 提出的 SeNet (Squeeze and Excitation Networks)，得到升级版的 YOLO 模型

SENet 中最关键的 SE 模块(见图 5)使用两个连续卷积层的输出并分流进入下一层，网络会将输入的

尺寸为 H * W * C 的特征图通过全局平均池化将每个特征图压缩 1 * 1 * C 的实例数列，这一步称为 Squeeze，然后再经过两个全连接层去构建通道之间的相关性，这一步称为 Excitation，接着通过 Sigmoid 函数将数据归一化到 0~1 的权重，最后通过 Scale 操作将权重加到每个通道上。通过两个全连接层的目的是限制模型的复杂度，同时增加泛化能力。而在两个全连接层中穿插一个 ReLu 激活函数是为了增加非线性因素从而是使整个模型更好的拟合。

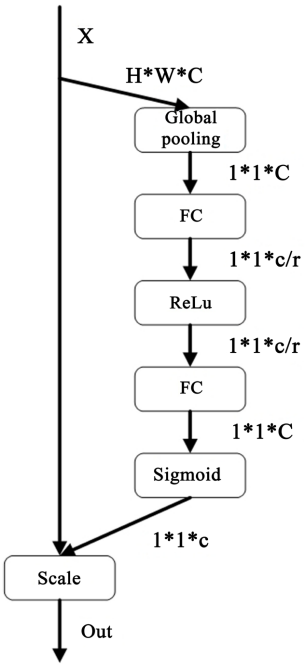

Figure 5. SE module
图 5. SE 模块

3.3. 改进的 YOLO 网络

ResNet (残差网络)是为了解决网络退化问题提出的，当神经网络进一步加深时，反向传播的梯度会非常小，对于深层的权重更新就十分困难。而残差网络提出在网络中添加直连通道，允许原始信息能够直接传入后面的层中，从而解决了梯度消失和梯度爆炸的问题。但是，他在我们的操作过程中也有不好的地方，当网络权重有一点轻微的变化时，就会引起较大的输出变化。如果采用 ResNet 结构，计算量将十分巨大。SENet 顺着空间维度进行特征的压缩，这种全局的操作可以提供一个很好的感受野，在多分类问题上有很好的泛化能力，并且我们通过在最后 Scale 操作时是通过逐通道加权到之前的特征上，完成通道维度对原始特征的重新标定。重新加权的操作，可以得到抑制无效的特征，并提升有效特征的权重，从而有效地提升网络性能，而计算量不会太大。

所以本文在原本 YOLO 网络的基础上用 SENet 网络(见表 1)对其特征提取网络进行了修改，并将产生不同的层输出的特征图输入特征金字塔结构中，加强网络的特征提取能力。

SEBottleneck 是一个组合，先进入两个 Conv + BN 层中初步提取特征，再进入 Conv + BN + ReLu 中提高其拟合能力。接着才进入 SE 模块中增加其全局视野的能力。不同的 Bottleneck 会经过不同次数的上述操作，最后获得特征图的通道数也不同，经过 SEBottleneck1 后的特征图尺寸缩小，通道数变为 256 通道，接着进入 SEBottleneck2 变为 512 通道，进入 SEBottleneck3 变为 1020 通道，进入 SEBottleneck4 变

为 1040 通道。改进后的 YOLO 网络会进入 RPN 层中通过 9 种大小不同的预测框对大小不同的障碍物进行检测，以满足我们实际设计的需要。

Table 1. SENet
表 1. SENet

	Type	Filters	Size	Output
	Convolutional	64	3*3	
	Convolutional	64	3*3	
	Convolutional	128	3*3	
	MaxPooling	128	3*3	
3x	Convolutional	128	1*1	
	Convolutional	256	3*3	
	Convolutional	256	1*1	
	se_module	256		
8x	Convolutional	256	1*1	
	Convolutional	512	3*3	
	Convolutional	512	1*1	
	se_module	512		
36x	Convolutional	512	1*1	
	Convolutional	1024	3*3	
	Convolutional	1024	1*1	
	se_module	1024		
3x	Convolutional	1024	1*1	
	Convolutional	2048	3*3	
	Convolutional	2048	1*1	
	se_module	2048		
	Avgpool		7*7	
	Droupout			
	Linear	1000		

本文最终的网络结果如图 6 所示。

4. 实验验证与分析

4.1. 数据集及其预处理

本文所使用的犬种数据集包括斯坦福大学搜集的 120 种犬类数据，以及 kaggle 竞赛所使用的有标签的训练数据，两个数据集包含的图片有部分重复。同时，用到了自己在网络搜集的图片以及通过人工打标的方式的部分数据。图片总和为 40,000 张。实验中随机抽 70%作为训练集，20%作为验证集。剩余 10%作为测试集。同时为了增强整个训练效果的鲁棒性，在进行训练前，对数据集做了随机镜像、适当高斯

噪声、垂直方向上图像旋转等处理。并对光线等进行了相应的随即处理，用于减少网络过拟合的同时，增加网络泛化的能力。之后对图片进行尺寸的统一处理，方便整个网络的输入。

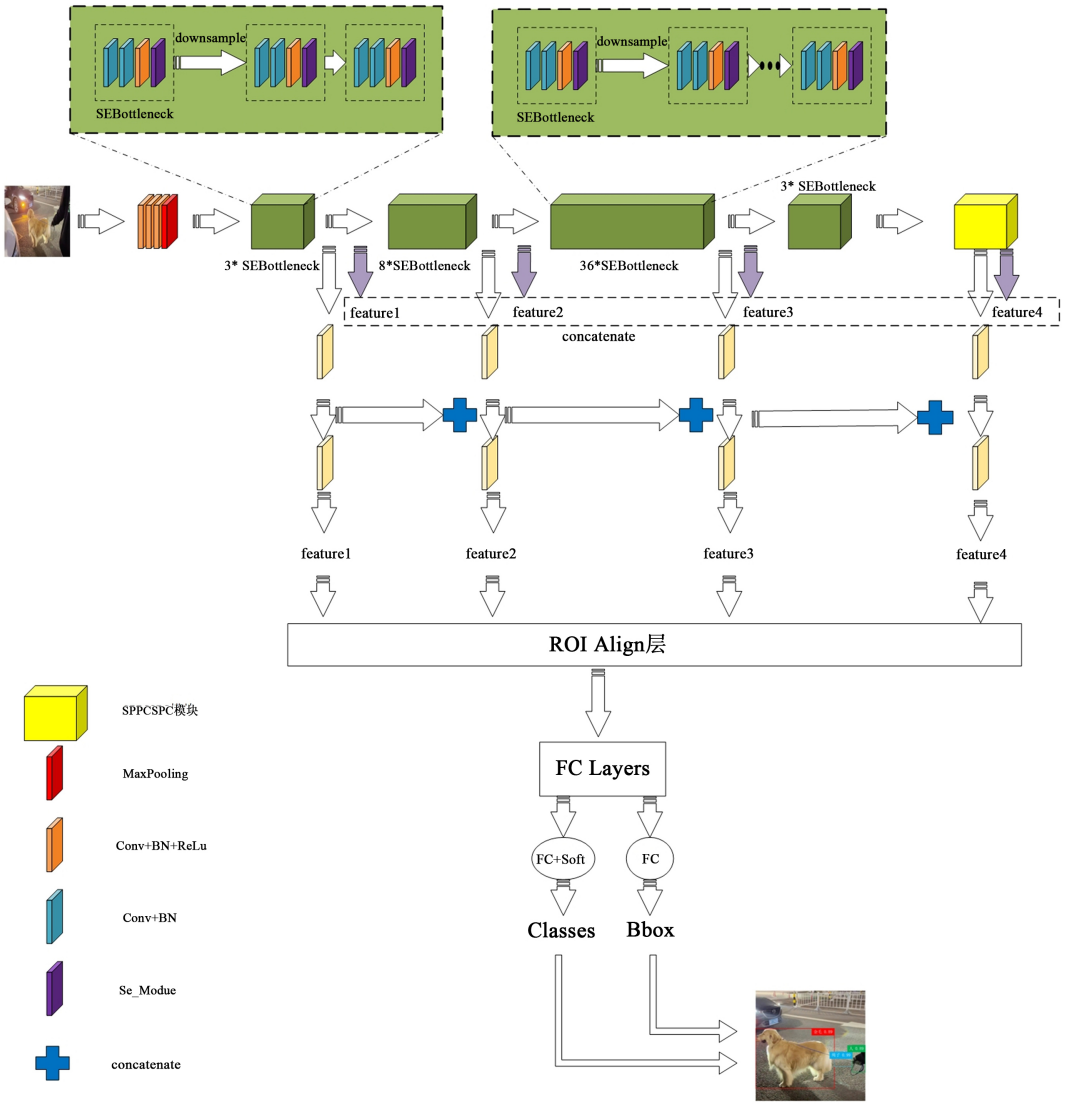

Figure 6. Complete network diagram
图 6. 完整网络示意图

4.2. 实验环境及评价指标

本实验采用硬件设备主要是 Intel(R) Xeon(R) W-2104 CPU @3.20 GHz 内存 64 G；NVIDIA GeForce GTX 1080Ti GPU。

模型的性能指标包括训练集、验证集、测试集的准确率以及训练集和验证集的损失率。其中训练集的置信度和损失率体现是整个模型训练的性能，验证集的损失率体现的是训练过程中是否出现了过拟合的现象。测试集的置信度则是直接体现了整个训练的模型的识别能力。这两个指标如下：

$$\text{Accuracy} = \frac{1}{M}\sum_{i=1}^{M} I\left(y_i = f\left(x_i\right)\right) \tag{1}$$

$$l_{grad}=\frac{1}{n}\sum_{i=1}^{n}\left(F\left(\nabla_x\left(e_i\right)\right)+f\left(\nabla_y\left(e_i\right)\right)\right) \tag{2}$$

其中，M 是样本数量；y_i 是标签是模型，$f\left(x_i\right)$ 预测结果，I 是条件判断函数；对于损失函数，我们采用常用的 l1 损失。但是我们希望误差太大时，不要产生对应的 loss，所以对于太大的 loss，经过 log 变换后就会变小。我们于是便使用这种处理离群点的方式。其中 $F\left(x\right)=\ln\left(x+\alpha\right)$，$\alpha\left(>0\right)$ 为超参数。

简单进行 log 变换会让损失函数对于边缘的误差变化不敏感。所以我们采用 Hu J 在 2019 年的论文[9]中提出的 l_{grad} 来互补边缘误差得到变化的敏感。

图 7 为 300 个 epoch 的 loss 变化图

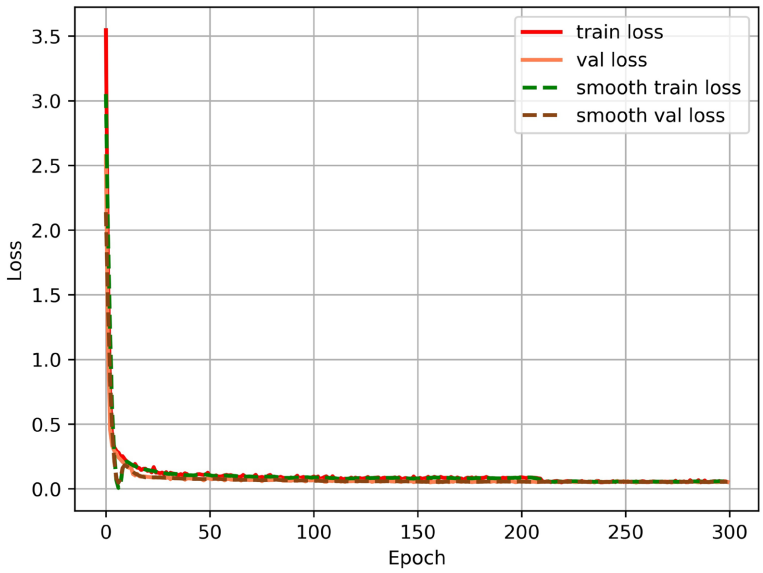

Figure 7. Loss change curve
图 7. 损失变化曲线

4.3. 结果分析

我们将如今主流的识别算法和本文优化后的网络进行了对应的比较，结果如表 2 所示：

Table 2. Comparison between mainstream recognition algorithm and algorithm experiment in this paper
表 2. 主流识别算法与本文算法实验对比

方法来源	所用方法	识别准确率/%
文献[6]	Alexnet 微调	63
文献[7]	拓展数据 + DCNN	67.31
文献[8]	Googlenet 迁移学习	69.07
本文	优化后的 YOLO 模型	68.05

由表 1 可以看出：文献[6]是基于传统的深度学习网络 Alexnet 而展开的训练，其检测精度明显低于其他方法，这是由于传统的网络参数训练队数据有着极大的依赖性，图像特征的提取也与先验知识有密切关系。特别是随着网络层数的不断增加，Alexnet 包含 8 个隐藏层，而本文所使用的网络层数明显更深，

单纯利用数据对网络进行训练是不能更好的实现分类任务的。文献[7]使用的数据多大 163 k 张，使用的训练方法是对 DCNNs 进行完全初始化训练。该方法虽然增加了数据总量，但训练学习到的特征过于单一。文献[8]的结果虽然比本文方法稍微高一点，不过在该实验中使用了比本文更多的数据集用于训练，并且在网络训练过程中特地训练一个狗脸探测器用于宠物识别对于本文的目的来说确实没有必要。

最终效果图如图 8 和图 9 所示：

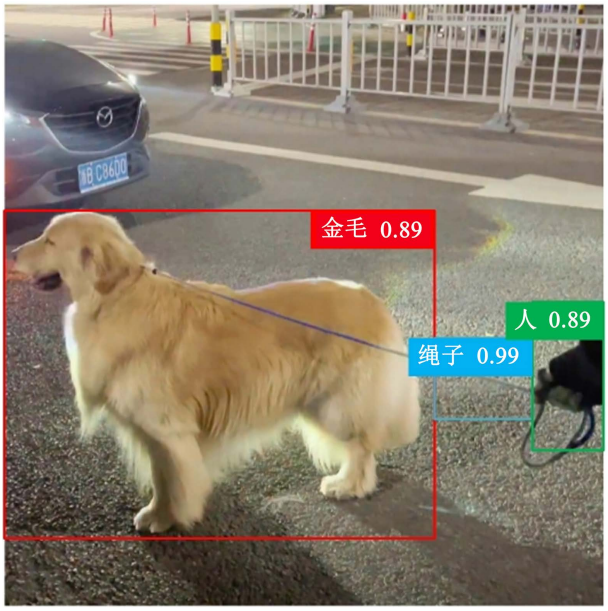

Figure 8. Model identification results
图 8. 模型识别结果

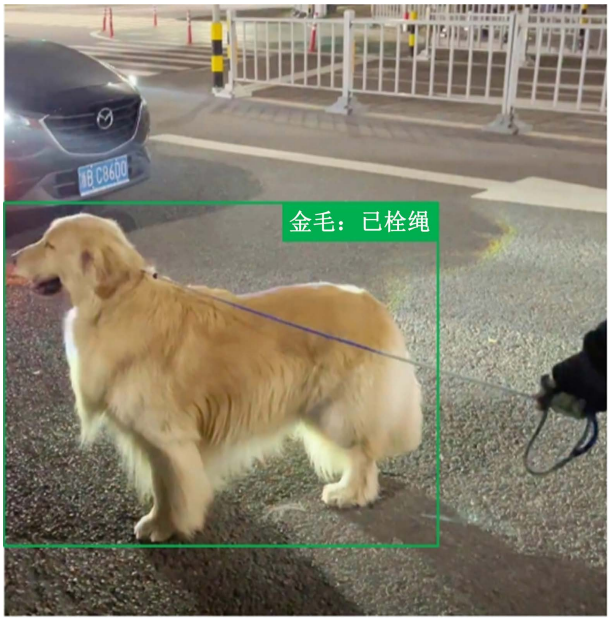

Figure 9. Final determination of illegal identification results
图 9. 非法识别结果最终判定

最后值得一提的是，在本文所搜集的公共数据集中存在部分犬类人工标注分类错误的情况。同时数据集的犬类数量并不均匀，这都是对于网络训练造成了一定的影响。

5. 结束语

本文提出了一套基于改进型的 YOLO 模型实现违法溜宠物现象识别。实验证明该方法对于属于细粒度图像分类的犬种有一定的性能改进。且针对实际需要，本文选取了 SENet 而非常用的 ResNet 特征提取网络来实现基础的特征提取，在准确率变动不大的情况下提升了训练速度。同时在特征提取网络最后阶段加入 SPPCSPC 模块扩大了整个感受野的更好地契合了最终实验结果的需要，这个重新构成的新网络是本文最大的创新点。

为了契合识别的需求，本文选用新的损失计算方式，针对性地提升了整个网络识别的效果。在精度变化不大情况下更好地收敛整个模型，此为创新点之二。最终通过实验表明整个网络能够很好地胜任生活中识别非法溜宠物现象的使用。

参考文献

[1] He, K., Gkioxari, G., Dollár, P., *et al.* (2017) Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2961-2969. https://doi.org/10.1109/ICCV.2017.322

[2] Wang, C.Y., Bochkovskiy, A. and Liao, H. (2022) YOLOv7: Trainable Bag-of-Freebies Sets New State-of-The-Art for Real-Time Object Detectors. arXiv e-prints

[3] 李思瑶, 刘宇红, 张荣芬. 基于迁移学习与模型融合的犬种识别方法[J]. 智能计算机与应用, 2019, 9(6): 101-106.

[4] 蒋俊蕊, 魏延, 王晶仪, 张文泷, 张昆, 李媛媛. 基于人体时空骨架特征的图卷积行为识别算法[J]. 重庆师范大学学报(自然科学版), 2022, 39(4): 124-133.

[5] 娄英欣. 基于深度学习的目标检测[D]: [博士学位论文]. 北京: 北京邮电大学, 2018.

[6] Wang, X., Ly, V., Sorensen, S., *et al*. (2015) Dog Breed Classification via Landmarks. 2014 *IEEE International Conference on Image Processing* (*ICIP*), Paris, 27-30 October 2014, 5237-5241. https://doi.org/10.1109/ICIP.2014.7026060

[7] 吴迪, 刘秀磊, 侯凌燕, 等. 基于显著性检测和迁移学习的花卉图像分类[J]. 北京信息科技大学学报: 自然科学版, 2019(1): 9.

[8] 张泽中, 高敬阳, 吕纲, 等. 基于深度学习的胃癌病理图像分类方法[J]. 计算机科学, 2018, 45(B11): 6.

[9] Hu, J., Ozay, M., Zhang, Y., *et al*. (2019) Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps with Accurate Object Boundaries. 2019 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, HI, 7-11 January 2019, 1043-1051. https://doi.org/10.1109/WACV.2019.00116