doi: 10.7690/bgzdh.2023.05.012

基于改进 Focus 的小目标检测技术

谢志宏，陈晓明
（陆军装甲兵学院兵器与控制系，北京 100072）

摘要：为解决深度神经网络模型小目标检测效果不佳的问题，对 Focus 结构进行改进，提出一种即插即用的轻量级结构 Focus+。搜集相关图像并建立包含 5 类目标的军事小目标数据集，将 Focus+插入常用的目标检测模型，使用不同尺度的输入进行了多组对比实验。实验结果表明：引入 Focus+模块后，模型检测的平均精度均值平均提高了 0.8 %，说明其能够在占用较少算力的同时提取到浅层网络的高分辨率特征，有效提高小目标的检测精度。

关键词：小目标检测；深度学习；特征提取；军事目标
中图分类号：TJ06　　**文献标志码：**A

Small Target Detection Technology Based on Improved Focus

Xie Zhihong, Chen Xiaoming
(*Department of Weapons and Control*, *Army Armored Force Academy*, *Beijing* 100072, *China*)

Abstract: In order to solve the problem of poor performance of deep neural network model in small target detection, the Focus structure is improved, and a plug-and-play lightweight structure Focus + is proposed. Relevant images are collected and a military small target dataset containing 5 types of targets is established. Focus + is inserted into the commonly used target detection model, and several groups of comparative experiments are carried out using different scale inputs. The experimental results show that the average accuracy of model detection is improved by 0. 8% after the introduction of Focus + module, which indicates that it can extract the high-resolution features of shallow network while occupying less computing power, and effectively improve the detection accuracy of small targets.

Keywords: small target detection; deep learning; feature extraction; military target

0　引言

随着战争形态由机械化向信息化、智能化转变，各式智能化武器装备层出不穷，引入高性能的目标检测算法，使其能够精准、高效地识别各类目标是研发智能化武器装备的重点和难点。以 Viola-Jones Detector、DPM[1]等算法为代表的传统方法特征提取能力有限，且检测速度与精度也不甚理想，难以适应复杂的真实战场环境。近年来，基于神经网络的目标检测算法不断发展，学者们提出了以 R-CNN[2-8]系列为代表的两阶段(two-stage)检测模型和以 YOLO[9-12]、SSD[13-17]系列为代表的一阶段(one-stage)检测模型，这些基于深度卷积神经网络算法的性能普遍更加优秀，已经成为目标检测的主流手段。

现有的算法对于中目标和大目标的检测已经取得了较好的效果。然而，在成像制导和遥感侦察这类以小目标检测(检测对象的宽、高均小于原图宽、高的 1/10)为主的应用场景中，很多模型的表现却不尽如人意。造成该现象的一个主要原因是样本图像中检测对象的面积占比很小，而背景噪声占比却很大，训练模型时，通常会将输入图像的分辨率调整至固定大小，使得检测目标的面积被进一步压缩，从而导致目标信息缺失。当数据传入模型一定深度时，经过几次下采样操作后，特征图中的目标基本上只有个位数的像素，模型难以学习到高分辨率特征，很容易丢失目标，自然不会取得太好的训练效果。解决这一问题最直接的方法就是设置高分辨率的输入，或是改为训练深度较浅的模型。具体来讲，直接使用高分辨率的输入会增加计算量，对于硬件不友好，且训练速度和检测速度均会明显降低，而绝大多数军事类应用场景都非常注重检测速度，故这样做并不妥当。另一方面，使用浅网络虽然能够带来较小的感受野，但模型结构的设计以及参数数量降低导致的性能下降都是非常棘手的问题。

基于上述背景，为了提升现有深度神经网络模型对于小目标检测的性能，笔者对 Focus 模块进行改进，提出一种即插即用的轻量级模块 Focus+。在显著提高输入分辨率的基础上通过切片操作进行下

收稿日期：2023-01-06；**修回日期：**2023-02-10
基金项目：军队重点项目(LJ20191A030128)
作者简介：谢志宏(1970—)，男，湖南人，博士，副教授，从事模式识别与计算机视觉研究。E-mail: 214223211@qq.com。

采样，再利用低参数量的分组卷积和逐点卷积来提取图像的特征，使模型更加注重浅层高分辨率特征的学习，并且能够输出 3 通道的特征图以适配各种现有模型的输入。

1　Focus+模块结构

Focus 模块诞生于 2020 年提出的 YOLOv5 网络。它的主体部分可以看作一个切片操作，采取隔行采样拼接的策略把高分辨率的输入图像拆分为多个低分辨率的图像，其本质是一种特殊的下采样操作。Focus 模块的具体结构如图 1 所示，若输入图像的尺寸为 320×320×3，则传入 Focus 时先复制 4 份，然后通过切片操作将其切成 4 张 160×160×3 尺寸的图像，随后在通道维度上拼接特征图，生成大小为 160×160×12 的张量，并进行一次卷积，最终通过批标准化与 SiLU 激活操作后将结果输入下一层。

笔者提出的 Focus+模块继承了 Focus 的切片操作，并在此基础上做出了其他改进。Focus+模块的具体结构如图 2 所示，首先将尺寸为 h×w×3 的输入切拼为 h/2×w/2×12 的特征图，把 h−w 平面上的信息转换至通道维度后再分别通过逐点卷积层和分组卷积层进行特征提取。这种结构能够有效抑制下采样所造成的信息损失。

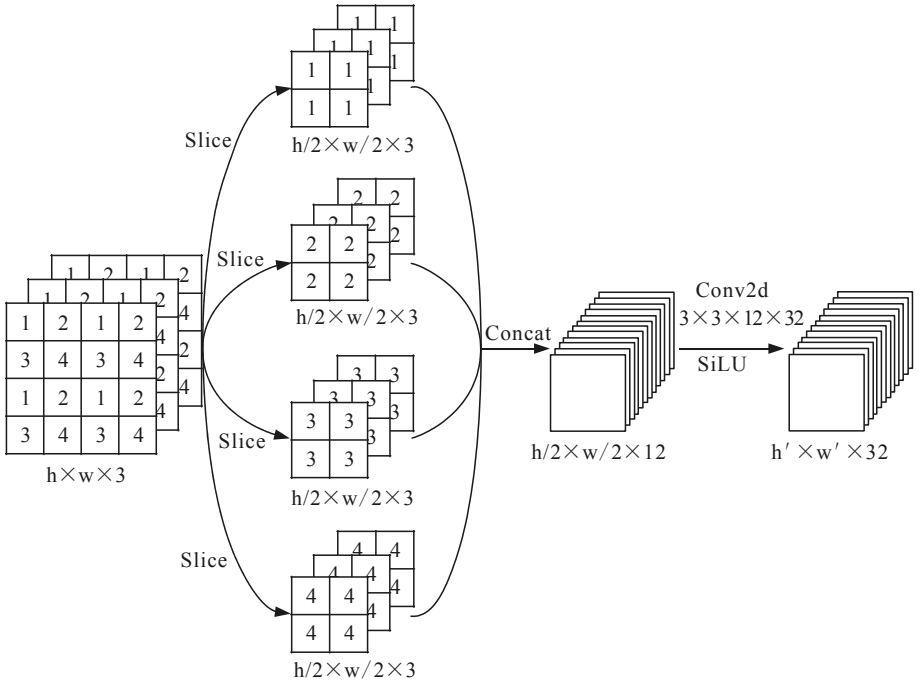


图 1　Focus 模块结构

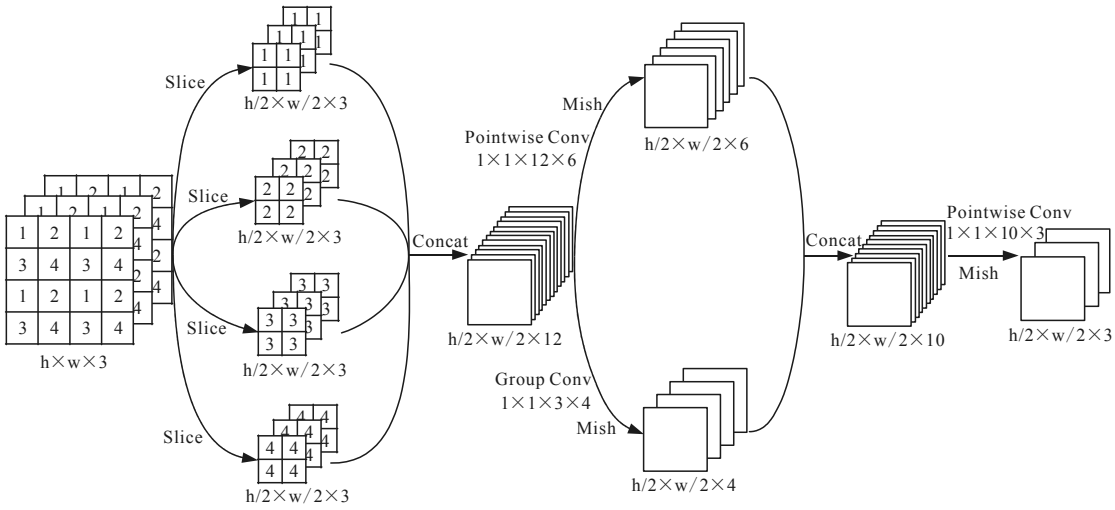


图 2　Focus+模块结构

由于所用逐点卷积的卷积核尺寸只有 1×1，使用它进行降维只会引入非常少量的参数。此外，模

型中的分组卷积核尺寸为 1×1×3×4，其参数数量相较于普通的 2 维卷积缩减了 75 %。将特征提取得到的结果在通道维度上进行融合，之后再传入一个逐点卷积层，将 Focus+的输出调整为 3 通道张量以适配其他模型图像的输入尺寸，从而增强模块的通用性。对于激活函数的使用，这里不再使用原有的 ReLU[18]系列，而是引入 Mish[19]进行激活，其表达式为：

$$f(x)=x\tanh(\ln(1+e^{x}))\text{。} \tag{1}$$

Mish 的函数图像如图 3 所示，其优势在于正值可以达到任何高度，避免了由于封顶而导致的饱和。对于负值的轻微容忍允许更好的梯度流，且平滑的激活函数能够使样本特征信息深入神经网络，从而获取更强的准确性和泛化性。

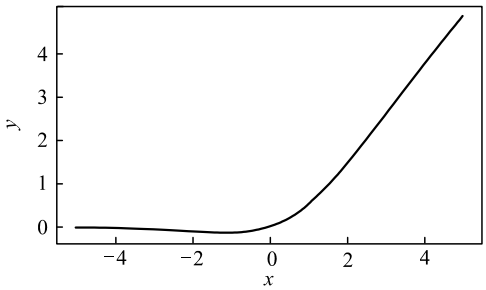


图 3 Mish 函数曲线

总的来说，Focus+是一种专注于图像中小目标检测的轻量化模块，旨在占用较少算力的同时提取更多的浅层高分辨率图像特征，并且很容易接入现有模型的输入端，能够做到即插即用，使用起来非常方便。

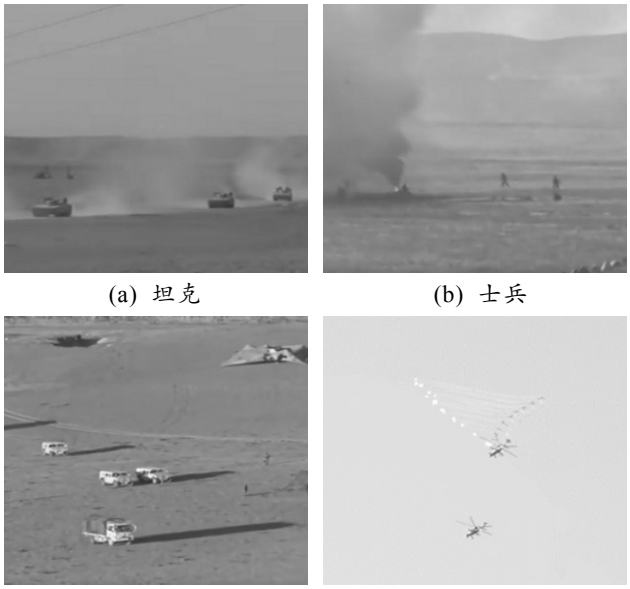
2 实验

实验基于 Windows 10 系统，PyTorch1.8.1 框架，硬件配置为 3.60 GHz Intel Core i7−9700K CPU，16 GB 内存，Nvidia GeForce RTX 2080 单 GPU 加速模型训练。

2.1 数据集

由于没有找到适用于本研究的公共集，因此自行收集样本并建立实验所用数据集，数据集包括：1) 通过互联网爬取到的相关视频(截图)与图像；2) 使用高清相机采集到的相关视频(截图)与图像。

经筛选和整理后得到了包含士兵、坦克、卡车、吉普车、直升机这 5 类目标共 8 370 张图像的数据集。数据集中的部分图像如图 4 所示，其中小目标占比 90%以上，但也存在少量像图 4(c)底部的卡车这样略大的目标。将样本图像按比例 resize 至 2 048×2 048 大小，对于少数尺寸较小的图像(如 milair−dataset 中 1 280×720 分辨率的图像)不进行拉伸，而是使用灰色(128, 128, 128)来填充背景。利用 LabelImg 软件进行标注，并按照 7:2:1 的比例划分训练集、验证集与测试集。



(a) 坦克 (b) 士兵

(c) 卡车、吉普车与士兵 (d) 直升机

图 4 数据集样本展示

2.2 结果与分析

为验证 Focus+的性能，将其插入常用网络进行对比实验。为确保实验的公平性，各组训练参数尽量保持一致，其中 batchsize 设置为 4，Darknet 的 batch/subdivisions 设置为 64/16，learning rate 取 0.001，epoch 设置为 250，各模型均采用默认参数的 Adam 优化。

基于主干网络为 Darknet−53 的标准 YOLOv3[11]模型进行测试，得到的训练结果如表 1 所示。表中所示的 mAP(mean average precision)指 COCO 数据集评价指标中的 AP，即 IoU 取 0.5 至 0.95 步长为 0.05 时 mAP 的均值，下同。根据表 1 中实验结果可知，当训练集样本的输入分辨率逐渐增加时，模型的 mAP 随之上升，检测速度有所下降。这是比较正常的现象，理论上增加输入图像的分辨率有助于模型学习到多尺度的特征，提高检测性能；同时，模型中的卷积、池化层等结构所产生的计算量会使检测速度大幅降低。在较高分辨率的同等输入条件下，引入 Focus+结构的模型精度明显有所提升，mAP 分别提升了 0.4%和 0.8%，而 FPS 却没有大幅变化，这说明专注于浅层特征进行学习是有效的。该模块对于 YOLOv3 而言能够提升检测精度，并且

不会占用太多算力。值得注意的是，320×320 输入下的模型 mAP 在插入了 Focus+之后下降了 1.4%，这是输入的分辨率过小导致的。在主干网络不变的情况下，经切片后特征图的尺寸显著减小了，而感受野却相对偏大，经过几次卷积之后目标的信息大量丢失，从而引起模型精度的降低。

表 1　YOLOv3 测试结果

Method	Backbone	Resolution	mAP	mAP_{50}	mAP_{75}	FPS
YOLOv3	Darknet53	320^2	15.2	43.1	10.0	25.13
YOLOv3	Darknet53	416^2	17.6	45.4	12.6	22.73
YOLOv3	Darknet53	608^2	20.3	47.7	15.3	12.69
YOLOv3	Darknet53	832^2	21.9	49.3	16.7	8.70
YOLOv3	Darknet53	$1\,216^2$	24.0	51.7	18.5	5.13
YOLOv3+ Focus+	Darknet53	320^2	13.8	41.5	8.7	23.96
YOLOv3+ Focus+	Darknet53	832^2	22.3	49.9	17.1	8.13
YOLOv3+ Focus+	Darknet53	$1\,216^2$	24.8	52.5	19.4	4.59

对于分支众多的 SSD 系列，考虑到在原始的 SSD300 网络上训练小目标数据集的难度太高了，准确率可能会很低，难以对比其结果，笔者选择使用 ResNet101[20]作为主干的 SSD[13]模型与强化小物体检测的 DSSD[15]改进模型进行实验，测试结果如表 2 所示。根据表 2 结果不难发现，当 SSD513 输入的分辨率提高至 1026×1026 时，mAP 不仅没有上升反而降低了 0.1 %。此前进行相关实验时，我们常认为不断提高输入分辨率即可使检测的准确率上升，而这样做只需考虑模型收敛以及检测速度方面的成本。这一现象的出现说明各模型对于输入都是有一定接受度的，当输入大小超出了其承受阈值时，其性能可能反而会下降。在引入 Focus+之后，1026×1026 输入下 SSD 网络的 mAP 上涨了 0.8 %，DSSD 的 mAP 上涨了 1.3 %，这表明 Focus+有助于模型接受更高分辨率的图像。

表 2　SSD 测试结果

Method	Backbone	Resolution	mAP	mAP_{50}	mAP_{75}	FPS
SSD513	ResNet101	513^2	13.9	42.0	10.1	5.59
DSSD513	ResNet101	513^2	16.2	45.9	13.0	4.46
SSD513	ResNet101	$1\,026^2$	13.8	42.0	10.1	1.36
DSSD513	ResNet101	$1\,026^2$	19.1	47.5	15.2	1.10
SSD513+ Focus+	ResNet101	$1\,026^2$	14.6	42.7	10.7	1.34
DSSD513+ Focus+	ResNet101	$1\,026^2$	20.4	48.7	16.3	1.11

同理，使用不同主干网络的 RetinaNet 进行测试，其结果如表 3 所示。根据表 3 不难发现，在引入 Focus+后模型的 mAP 有所提升，分别提高了 1.5%和 0.5%，并且 FPS 的变化仍旧不是很明显，再次验证了其对于模型小目标检测性能提升的有效性。

表 3　RetinaNet 测试结果

Method	Backbone	Resolution	mAP	mAP_{50}	mAP_{75}	FPS
RetinaNet	ResNet50	500^2	18.9	42.4	15.1	***8.88***
RetinaNet	ResNet101	500^2	21.7	45.1	18.1	7.25
RetinaNet	ResNet50	$1\,000^2$	21.3	44.5	17.8	2.29
RetinaNet	ResNet101	$1\,000^2$	24.6	47.8	20.9	1.86
RetinaNet+Focus+	ResNet50	$1\,000^2$	22.8	46.1	19.0	2.21
RetinaNet+Focus+	ResNet101	$1\,000^2$	25.1	48.6	21.3	1.79

3　结论

笔者主要针对小目标检测问题进行研究，建立相关数据集，并提出一种即插即用的轻量化模块 Focus+，在保证检测速度的前提下，能提升深度神经网络对于军事小目标的检测精度，得出的主要结论如下：

1) 引入 Focus+模块能够有效提升深度神经网络模型的小目标检测效果，实验中模型的平均精度均值平均可以提高 0.8%，并且对于检测速度的影响不大。

2) 当传入网络的图像分辨率过低时，引入 Focus+会导致模型的检测性能下降，其具体的阈值因模型而异。

3) 当传入网络的图像分辨率逐渐增加时，模型的检测精度也随之上升，但过高分辨率的输入反而会导致检测效果降低，在模型中插入 Focus+模块能够缓解这种退化现象。

后续工作可以考虑改进 Focus+的特征提取部分结构，包括卷积层的类型、输出通道数以及连接方式等，通过尝试多种不同的组合并进行大量实验来确定最佳方案，以进一步提高其性能。

参考文献：

[1] RANJAN R, PATEL V, CHELLAPPA R. A deep pyramid deformable part model for face detection [C]//Proceedings of the 7th IEEE International Conference on Biometrics Theory, Applications and Systems. Washington D. C. USA: IEEE Press, 2015: 1−8.

[2] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Columbus, USA, 2014: 580−587.

[3] GIRSHICK R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. Boston, USA, 2015: 1440−1448.

[4] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137−1149.

[5] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.

[6] LIN T, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii, USA, 2017: 2117−2125.

[7] HUANG J, RATHOD V, SUN C, et al. Speed/accuracy trade- offs for modern convolutional object detectors[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Piscataway: IEEE, 2017: 3296−3297.

[8] SHRIVASTAVA A, SUKTHANKAR R, MALIK J, et al. Beyond Skip Connections: Top-Down Modulation for Object Detection[J]. arXiv preprint arXiv: 1612. 06851, 2016.

[9] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. LosAlamitos: IEEE Computer Society Press, 2016: 779−788.

[10] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, 2017: 7263−7271.

[11] REDMON J, FARHADI A. Yolov3: An incremental improvement[C]//Computer Vision and Pattern Recognition. Berlin/Heidelberg, Germany: Springer, 2018: 1804−2767.

[12] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv: 2004.10934, 2020.

[13] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Springer, Cham. Amsterdam, Netherlands, 2016.

[14] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834−848.

[15] Fu C Y, Liu W, Ranga A, et al. Dssd: Deconvolutional single shot detector[J]. arXiv preprint arXiv: 1701. 06659, 2017.

[16] Li Z, Zhou F. FSSD: feature fusion single shot multibox detector[J]. arXiv preprint arXiv:1712.00960, 2017.

[17] SHEN Z, ZHUANG L, LI J, et al. DSOD: Learning Deeply Supervised Object Detectors from Scratch[C]// 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017.

[18] GLOROT X, BORDES A, BENGIO Y. Deep Sparse Rectifier Neural Networks[J]. Journal of Machine Learning Research, 2011, 15: 315−323.

[19] MISRA D. Mish: A Self Regularized Non-Monotonic Neural Activation Function[C]// 2019 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Long Beach, USA, 2019: 847−861.

[20] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.