Computer Science and Application 计算机科学与应用, 2021, 11(4), 1061-1069
Published Online April 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.114109

Hans汉斯

基于改进SSD算法的小目标检测与应用

刘　洋，战荫伟

广东工业大学计算机学院，广东　广州
Email: 476028107@qq.com

收稿日期：2021年3月23日；录用日期：2021年4月19日；发布日期：2021年4月27日

摘　要

针对通用目标检测方法在复杂环境下检测小目标时效果不佳、漏检率高等问题，本文对SSD小目标检测算法进行改进。利用训练损失的反馈作为判断条件，结合数据增强提高模型对复杂环境的抗干扰能力，降低小目标的漏检率，在网络中引入注意力机制，增加SENet (Squeeze-and-Excitation)模块，对模型中的特征通道进行权重重分配，对无效的特征权重进行抑制，提升有用的特征权重占比。实验结果表明，相比原SSD算法，改进的SSD算法在不引入过多计算量的情况下，能够有效弥补训练过程中小目标监督不到位的不足，在VOC数据集和工地安全帽佩戴数据集上，精度都得到了明显提升。

关键词

SSD，深度学习，小目标检测

Small Object Detection and Application Based on Improved SSD Algorithm

Yang Liu, Yinwei Zhan

School of Computers, Guangdong University of Technology, Guangzhou Guangdong
Email: 476028107@qq.com

Received: Mar. 23rd, 2021; accepted: Apr. 19th, 2021; published: Apr. 27th, 2021

Abstract

To address the problems of ineffective detection of small objects and high miss detection rate of generic object detection methods in complex environments, this paper is to improve SSD small object detection algorithm. In order to avoid inadequate supervision of small objects in the training process, the feedback of training loss is used as the judgment condition, combined with data enhancement to improve the anti-interference ability of the model in complex environments, re-

文章引用: 刘洋, 战荫伟. 基于改进 SSD 算法的小目标检测与应用[J]. 计算机科学与应用, 2021, 11(4): 1061-1069.
DOI: 10.12677/csa.2021.114109

ducing the miss detection rate of small targets. The attention mechanism is introduced in the network and the SENet (Squeeze-and-Excitation) module is added to redistribute the weights of the feature channels in the model, which suppress the invalid feature weights and increase the percentage of useful feature weights. The experimental results show that compared with the original SSD algorithm, the improved SSD algorithm significantly improves the detection accuracy on both the VOC dataset and the safety helmet wearing dataset without too much computational effort.

Keywords

SSD, Deep Learning, Small Object Detection

1. 引言

目标检测是计算机视觉中最具有挑战性的任务，目的是在图像的复杂背景下找到若干目标，并对每一个目标给出一个精确的目标包围盒并判断包围盒中的目标所属的类别[1]。深度学习的兴起使得目标检测得到加速发展，准确性和实时性都得到了提升，如Girshick等人提出的R-CNN、Fast R-CNN算法[2] [3]、Ren等人提出的Faster R-CNN算法[4]、Joseph等人提出的YOLO算法[5]以及Liu等人提出的SSD (Single Shot MultiBox Detector)算法[6]等。其中，小目标检测是目标检测领域中一个重要的难点问题，实际应用场景复杂、小目标信息不充分，导致小目标的检测效果始终不是很好。小目标检测因而成为计算机视觉领域中的一项具有巨大挑战性的任务。

上述方法仅对常规的目标检测问题效果较好，但所提取出的特征对小目标的表示能力较差，检测效果不佳。MS COCO数据集中将尺寸小于32 × 32像素的目标定义为小目标[7]，大于32 × 32像素小于96 × 96像素的目标定义为中目标，大于96 × 96像素的目标为大目标。Huang等人[8]对现阶段的检测器进行调研发现，现阶段的目标检测系统的精度，在小目标上的精度普遍比大目标低10倍，原因主要是由于样本中的小目标分辨率太低，虽然卷积神经网络的特征提取能力对于大中目标已经足够，但是对于小目标还是力不从心，小目标能提供给模型的信息过少也是制约目标检测发展的瓶颈之一。对此，一系列针对小目标检测的方法应运而生，小目标检测因而成为热点研究领域。

Fu等人[9]提出DSSD算法，利用ResNet [10]替换SSD中的VGG [11]模型，同时为了减少小目标的漏检率，加入反卷积层(Deconvolution)，将图像分为更小的格子，但因为ResNet中引入残差连接等，算法的额外开销较大，比SSD算法的速度略慢。Singh等人[12]从训练角度切入，在数据层面思考，对数据集进行分析，发现训练样本中的小目标在待检测的图像中占比较小，于是采用一种多尺度的训练方式——图像的尺度归一化(SNIP)，在金字塔模型的每一个尺度上进行训练，高效利用训练数据，检测效果得到显著提升，但是计算成本巨大。Lin 等人[13]利用特征金字塔网络(FPN)融合模型高低层语义信息，增强模型提取的特征对小目标的表达能力。虽然上述方法都在一定程度上提升了小目标的检测精度，由于网络模型冗余导致的算法实时性不足、模型轻量化导致的精度不够、数据量不平衡导致训练不充分等因素，上述方法在实际场景下的检测效果仍然不理想。

本文基于SSD方法，利用数据增强和注意力机制设计一种小目标检测算法，在增加计算量可近似忽略的前提下，提升检测精度。首先，对训练过程进行优化，采用数据增强的方法加强模型对小目标的监

督，以每次迭代过程中的各项目标损失占比为判断依据，确定是否在下次迭代过程中增强输入数据，若在当前迭代中小目标贡献的损失占比小于给定阈值，则下次迭代输入为增强图像，反之输入原图像；此外加入 SENet 模块，提升有效的特征信息权重，抑制作用较小的特征信息权重。实验结果表明，改进后的 SSD 算法优于原 SSD 算法，在实际场景下也能有很好的检测效果。

2. SSD 算法

2.1. 网络模型

SSD 的主要特点是在不同尺度上进行检测与识别，其网络模型分为基础网络和附加网络，在基础网络的末端添加了几个特征层作为附加网络用于预测不同尺度目标以及包围盒的偏移量和置信度。该算法以 VGG-16 模型作为特征提取网络并将其全连接层替换为卷积层，网络输入 RGB 三通道图像，附加网络附加 4 个卷积层。为提高目标检测精度，SSD 算法在不同的尺度上进行检测，如图 1 所示，图像输入网络后从左至右得到 6 层不同尺寸的特征图(feature map) Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2、Conv11_2，尺寸分别为 38 × 38、19 × 19、10 × 10、5 × 5、3 × 3、1 × 1，借鉴 Faster RCNN 算法中的锚点思想，在特征图上生成不同尺度不同宽高比的先验框，而后通过非极大抑制(NMS)等方法输出最终目标类别和定位结果。

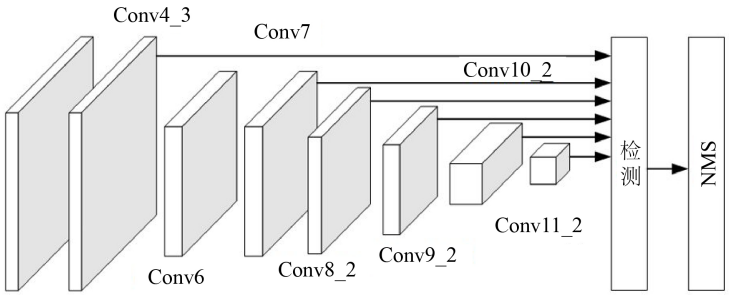

Figure 1. SSD network model
图 1. SSD 网络模型

2.2. SSD 优缺点分析

作为单步检测器，SSD 算法以回归的方式进行分类和定位，并结合 Faster RCNN 中的锚点思想进行预测，其算法中的先验框可以使得模型更快的收敛，降低训练成本。若没有采用锚点，则直接回归预测目标的坐标位置，模型难以收敛且计算成本巨大。同时 SSD 算法选取了模型中的六个特征图进行多尺度检测，各个卷积层所输出的特征图包含的信息是不同的，深层特征尺寸小，包含的语义信息更丰富，适合检测大目标，而浅层特征尺寸大，细节纹理特征信息更为丰富，对小目标的检测很有帮助，其采用的多尺度检测一定程度上是有利于提升小目标的检测精度的。

虽然 SSD 采用了多尺度的检测机制，利用 VGG-16 作为网络模型，通过 Conv4_3 的大尺度特征图来预测小目标，但是该层离顶层距离仍然较远，一个 32 × 32 的目标经过卷积后在 Conv4_3 特征图上大小仅为 4 × 4，如此少的像素信息难以对目标进行预测。另一方面，虽然采用了六个特征图进行预测，但是特征与特征之间都是独立的，实际上，模型的底层高分辨率特征由于经过的卷积运算较少，包含更多的纹理和细节信息，但包含的语义信息不足，难以区分目标和背景，而高层低分辨率特征语义信息丰富，但卷积下采样过程中丢失了大量细节信息。同时，浅层特征图中包含大量通道，不同通道对于网络的判断也有差别，有些通道包含的信息更为丰富，有些则不那么重要，SSD 算法并没有在通道上进行注意力

关注。综上，SSD 算法在实际应用场景下，对于小目标的检测效果并不理想。

3. 算法改进

3.1. 数据增强

Mosaic 方法[14]是由 Bochkovskiy 等人提出的一种数据增强方法，可以当作是 Cutmix [15]方法的改进版。Cutmix 算法对一张图像进行操作，在待增强图像上随机生成一个裁剪框，在裁剪框内填充训练集中其它数据中相应的位置像素，可以提高训练的效率。而 Mosaic 方法在 Cutmix 基础上再加入两张图像，如图 2 所示，采用 4 张图像进行混合，极大程度地丰富了训练图像的背景以及上下文信息，可以一定程度降低小目标的漏检率，同时 4 张图像的混合使得 mini-batch 不需要太大，可降低硬件需求。

Figure 2. Mosaic data enhancement example
图 2. Mosaic 数据增强示例

3.2. 训练优化

目标在现实生活中其实是随处可见的，比如远距离的交通标志，空中大背景下的小鸟等。MS COCO 数据集中设定小于 32 × 32 大小的目标即为小目标，小目标虽然是很常见的，但其分布却是不可预测的。MS COCO 数据集中，有 41.43%的目标是小目标，远远高于其他两种尺度的目标。然而，包含小目标的图像只占训练集所有图像的 51.82%，而包含大尺度目标和中等尺度目标的图像占比分别为 70.07%和 82.28%。也就是说，现有的环境下，待检测的大部分目标都是小目标，但是几乎一半的图像是不包含小目标的。这种训练样本的不平衡，也就导致训练的不平衡，严重阻碍了训练过程的推进。因此，考虑结合数据增强的策略从训练过程中去改善这种不平衡。模型在训练过程中大中小目标都是会反馈一定的损失进行优化驱动的，而在这一过程中，小目标所贡献的损失往往偏低，这样一来，训练好的模型对于大中目标的检测效果自然更好，这里提出一种优化方法，根据小目标所贡献的损失占总损失的比例来进行训练优化，如图 3 所示。假定某次迭代 d 中，来自小目标贡献的损失占总损失的比例小于给定阈值 μ(即表示这次迭代小目标受到的监督不足)，则第 $d+1$ 次迭代采用数据增强后的图像输入，反之，则采用原常规训练图像(此处阈值根据经验设置，经过多次不同阈值的设置对比，取 0.1 最为合适)。对于目标 o，其面积 S_o 可以近似于它的包围盒宽高之积 $ho \times wo$，S 表示第 d 次迭代中的小目标，MS COCO 数据集规定面积小于 32 × 32 的目标即为小目标，L_S^d 为第 d 次迭代下小目标的总损失，L^d 为第 d 次迭代下的总损

失，μ_s^d 为小目标的损失比。这种优化方法能够损失分布不均匀以及训练样本不平衡的问题。

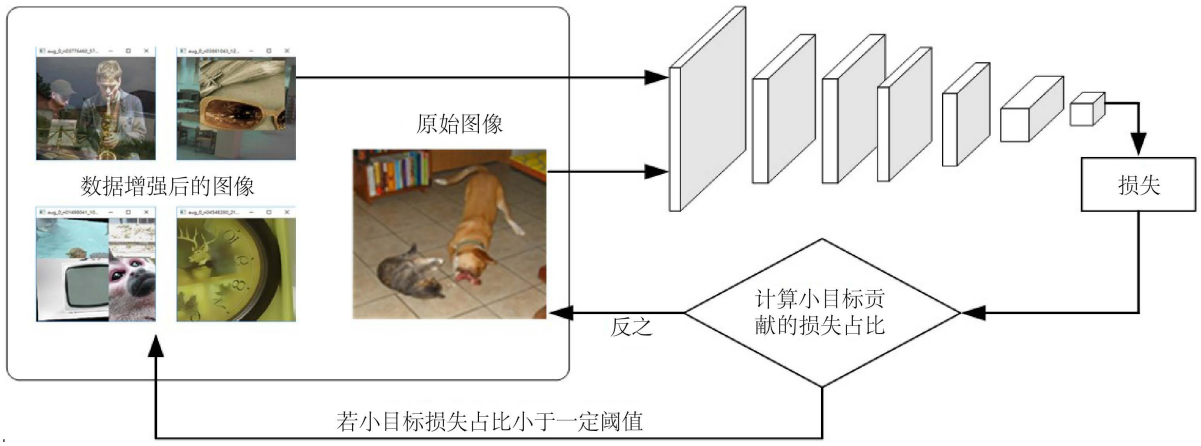

Figure 3. Training optimization pipeline
图 3. 训练优化流程图

3.3. 模型优化

SENet (Squeeze-and-Excitation Networks) [16]是 Hu 等人结合注意力机制提出的一种网络结构，Hu 等人曾凭借该结构夺得 ImageNet2017 竞赛图像分类任务冠军。该网络的核心思想在于，通过学习的方式，自动获取网络中每一个特征通道的重要程度，而网络也可以依据这个重要程度去提升有用的特征通道重要性并且抑制对于当前任务用处不大的通道重要性。输入特征 X，经过一系列卷积池化操作得到特征通道数为 C 的特征 U。首先对特征 U 进行压缩(Squeeze)操作，沿着空间维度进行特征压缩，得到通道级的全局特征，这个特征某种程度上具备全局感受野，且输出的维度与输入的特征通道数相匹配，表示在特征通道上响应的全局分布，可以让较浅的层获得更为全面的感受野。而后进行激励(Excitation)操作，学习各个通道之间的关系，通过参数 w 为每个特征通道生成权重。最后再进行权重的重新分配，完成在通道维度上对于特征的重标定，使得模型对于各个通道的特征更有判别能力。

因此，考虑到 SSD 算法模型中不同的通道对于小目标的检测也有着不同影响，也可以通过增强有效通道的特征权重，抑制无效通道的特征权重，从而提升 SSD 算法对于小目标检测的精度。所以选择在 Conv4_3、Conv7、Conv8_2、Conv9_2、Conv10_2、Conv11_2 层之后加入 SE 模块对特征图的特征通道进行权重重分配，改进后的结构图如图 4 所示。

4. 实验

实验平台采用如下配置：Ubuntu18.04 操作系统，基于 Pytorch1.3 框架搭建实验，GPU 显卡型号为 RTX2080ti，为充分展现实验效果，采用两套数据集，分别为 Pascal VOC 数据集以及 SHWD 数据集(工地场景安全帽数据集)，其中，为避免产生过拟合等由于数据量不充足造成的问题，选用 VOC2007 和 VOC2012 数据集的训练集作为训练集，在 VOC2012 数据集的测试集上进行测试。SHWD 数据集提供了用于安全帽佩戴和人头检测的数据，包括 7581 张图像，其中 9044 个佩戴安全帽的样本和 111514 个正常头部样本。

实验采用 VOC2007 数据集和 VOC2012 数据集进行训练，选取目前几个比较流行的目标检测方法进行对比，包括 YOLO 系列、SSD 系列、RCNN 系列，可以看出算法的平均精度得到了有效的提高，如表 1 所示。

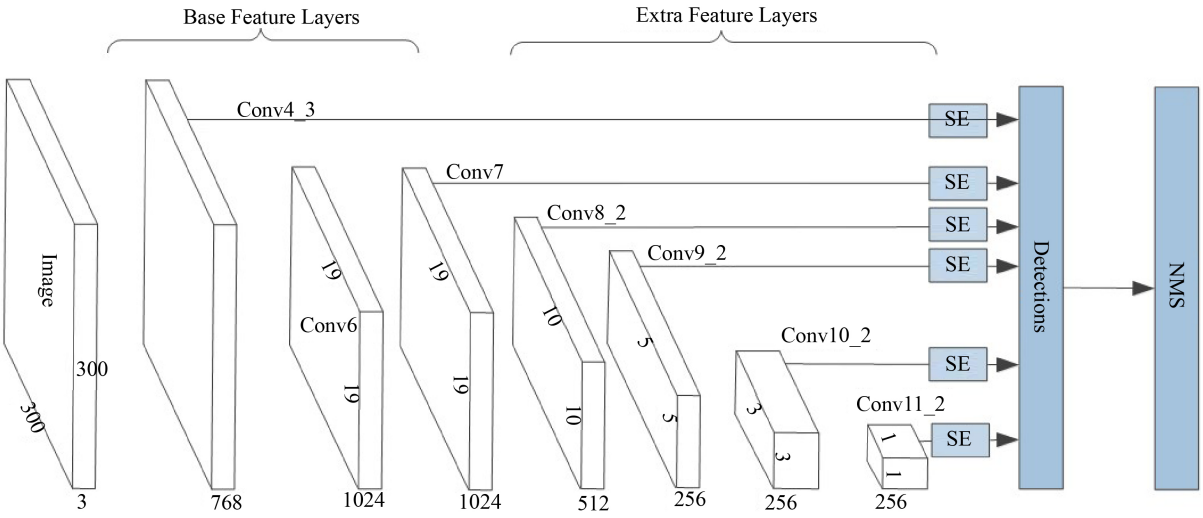

Figure 4. Improved network structure diagram
图 4. 改进后的网络结构图

Table 1. Comparison of mAP indexes of several target detection algorithms
表 1. 几种目标检测算法 mAP 指标对比

算法名称	数据集	Map (%)	骨干网络
FastRCNN	07 + 12	70.4	VGGNet
YOLOv2	07 + 12	73.8	DarkNet19
YOLOv3	07 + 12	77.9	DarkNet53
FasterRCNN	07 + 12	73.7	VGGNet
FasterRCNN	07 + 12	76.5	ResNet101
SSD300	07 + 12	77.6	VGGNet
SSD512	07 + 12	78.5	VGGNet
DSSD	07 + 12	78.6	ResNet101
Ours	07 + 12	80.6	VGGNet

具体各类的精度如表 2 所示，在 VOC2007 验证集上进行验证，将本文改进的 SSD 算法与原算法对比。可以得知，改进后的模型在 bird、bottle、plants、chair 等不同小目标类别上相比于原 SSD 算法均有着不同程度的提高。

Table 2. Comparison of different types of AP indexes
表 2. 不同类别的 AP 指标对比

算法名称	数据集	mAP(%)	aero	bike	bird	boat	bottle	bus	car	cat	chair
SSD	07 + 12	77.6	79.9	85.1	76.1	71.7	54.0	85.6	85.9	87.1	58.6
Ours	07 + 12	80.6	83.8	87.7	79.1	75.2	59.2	88.9	87.7	88.3	64.2
	cow	table	dog	horse	motorbike	person	plant	ship	sofa	train	tv
SSD	83.6	76.3	85.2	87.3	86.2	79.1	50.3	78.9	77.6	87.5	77.1
Ours	85.0	77.6	86.4	88.6	86.8	81.9	57.9	81.7	82.1	88.7	81.2

图 5 是原 SSD 算法与本文改进算法的效果对比，可以看出，原 SSD 算法对于 chair、bottle、plant 等小尺寸目标检测效果较差，漏检率较高，而改进后的 SSD 算法一定程度上降低了小目标的漏检率，并且提升了检测小目标的精度。

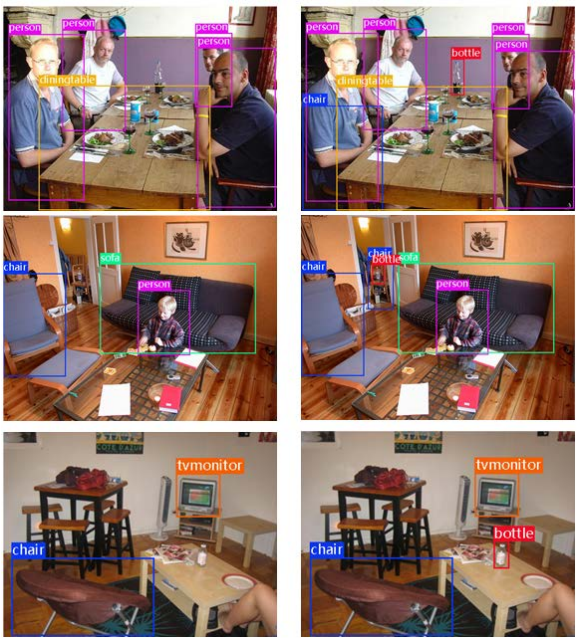

Figure 5. Comparison of SSD algorithm and improved SSD algorithm
图 5. SSD 算法效果与改进 SSD 算法效果对比

除了在标准公共数据集下进行比较，为了进一步探究算法的实际应用有效性，在相同的实验环境下，通过 SHWD 数据集中的安全帽佩戴进行训练，将训练完成后的模型进行测试，测试指标如表 3 所示，可以看出改进后的算法在该数据集上的平均精度是要高于 SSD 算法的，同时在识别安全帽是否佩戴的两种情况下，精度都是要高于原 SSD 算法的。如图 6 所示，图 6 左侧代表改进 SSD 算法的效果，图 6 右侧表示原 SSD 算法效果。由图 6 可以看出，工地场景下遮挡情况较多，原 SSD 算法容易因为遮挡而没有检测到相应目标，甚至可能因为环境光线等因素而出现误判，而改进 SSD 算法在遮挡情况下仍然成功检测到相应目标。

综合上述实验可得，本文提出的改进 SSD 算法相较改进前的算法，一定程度上降低了小目标的漏检率，同时精度得到有效提高，对于小目标以及遮挡目标的检测效果也更好，在实际的应用场景下如文中选用的工地环境下，也能得到充分应用，具有一定的实用价值。

Table 3. Comparison of two methods in construction site
表 3. 工地场景下两种方法对比

算法名称	AP (%)		mAP(%)
	佩戴安全帽	未佩戴安全帽	
SSD300	86.5	87.6	87.1
SSD512	87.7	88.5	88.1
Ours	90.6	91.1	90.9

Figure 6. Comparison of two methods in construction site
图 6. 工地场景下两种方法效果对比

5. 结语

本文针对小目标检测效果不佳、漏检率高等问题，对 SSD 算法进行改进，从小目标的训练损失占比切入，对每次迭代过程中的损失占比进行监督，结合一些数据增强方法增强了小目标的训练效果，在计算量增加成本可以忽略不计的情况下提高了检测效果。并在模型中引入 SENet 模块，筛选通道间的注意力并学习通道之间的相关性，对每层的特征通道进行权重重分配，最终改进的 SSD 算法在小目标的检测效果上得到了很大的改善，同时在安全帽佩戴数据集上表现优异，在实际场景下也具备一定的应用价值。未来将继续研究训练过程的优化和网络结构的调整以及从多尺度的方面入手，争求进一步提高精度。

参考文献

[1] Zou, Z., Shi, Z., Guo, Y. and Ye, J. (2019) Object Detection in 20 Years: A Survey. arXiv:1905.05055.

[2] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the* 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[3] Girshick, R. (2015) Fast R-CNN. *Proceedings of the* 2015 *IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[4] Ren, S., He, K., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. 28*th Conference on Neural Information Processing Systems*, Montreal, 8-13 December 2014, 91-99.

[5] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[6] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., *et al*. (2016) SSD: Single Shot Multibox Detector. *European Conference on Computer Vision*, Amsterdam, 8-16 October 2016, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[7] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al*. (2014) Microsoft Coco: Common Objects in Context. *European Conference on Computer Vision*, Zurich, 6-12 September 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[8] Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., *et al*. (2017) Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. *Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 3296-3297. https://doi.org/10.1109/CVPR.2017.351

[9] Fu, C.Y., Liu, W., Ranga, A., Tyagi, A. and Berg, A.C. (2017) DSSD: Deconvolutional Single Shot Detector. arXiv: 1701.06659.

[10] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the* 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778.

https://doi.org/10.1109/CVPR.2016.90

[11] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv: 1409.1556.

[12] Singh, B., Najibi, M. and Davis, L.S. (2018) SNIPER: Efficient Multi-Scale Training. 32*nd Conference on Neural Information Processing Systems*, Montreal, 3-8 December 2018, 9310-9320.

[13] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the* 2017 *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106

[14] Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M. (2020) YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv: 2004.10934.

[15] Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y. and Choe, J. (2019) CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *International Conference on Computer Vision*, Seoul, 27 October-2 November 2019, 6022-6031. https://doi.org/10.1109/ICCV.2019.00612

[16] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745