Geomatics Science and Technology 测绘科学技术, 2022, 10(2), 61-68
Published Online April 2022 in Hans. http://www.hanspub.org/journal/gst
https://doi.org/10.12677/gst.2022.102007

Hans汉斯

基于改进YOLOv3的红外影像目标识别算法研究

王安祺，梁祺策，黄　鹤

北京建筑大学，测绘与城市空间信息学院，北京

收稿日期：2022年3月11日；录用日期：2022年4月12日；发布日期：2022年4月20日

摘　要

针对于夜间自动驾驶目标检测行人和车辆目标准确率低的问题，本文提出一种基于改进YOLOv3的红外影像目标识别算法。首先，该算法在原有残差单元基础上进行了改进，同时增加backbone中大尺寸图像的卷积次数，提高特征提取能力，并将后续常规卷积更换为深度可分离卷积，降低模型参数量，提高运行速度；其次，将其多尺度特征融合中特征融合结构更换为Panet结构，提高底层信息的利用率；最后，采用Distance-IoU (DIoU)作为archor损失函数，加快模型收敛。在Flir影像数据集上的测试结果表明，所提改进的YOLOv3红外识别算法改进的模型在模型大小几乎不变的情况下在准确率和召回率上获得较好的检测精度，相比于YOLOv3在行人和汽车两类上分别有2.94%和3.12%提升，平均AP也有3.03%的提升。实验证明，本方法改进后在提高检测精度的同时，还减少了模型量，提高了检测速度。

关键词

目标识别，行人车辆检测，YOLOv3，Backbone，特征融合，损失函数

Research on Infrared Image Target Recognition Algorithm Based on Improved YOLOv3

Anqi Wang, Qice Liang, He Huang

School of Geomatics and Urban Spatial Informatics, Beijing University of Civil Engineering and Architecture, Beijing

Received: Mar. 11th, 2022; accepted: Apr. 12th, 2022; published: Apr. 20th, 2022

文章引用: 王安祺, 梁祺策, 黄鹤. 基于改进 YOLOv3 的红外影像目标识别算法研究[J]. 测绘科学技术, 2022, 10(2): 61-68. DOI: 10.12677/gst.2022.102007

Abstract

Aiming at the problem of low accuracy of pedestrian and vehicle target detection in automatic driving at night, this paper proposes an infrared image target recognition algorithm based on improved YOLOv3. First, the algorithm improves the feature extraction ability by increasing the number of convolutions of large-size images in the backbone, and replaces subsequent conventional convolutions with depthwise separable convolutions to reduce the amount of model parameters and improve the running speed; In the scale feature fusion, the feature fusion structure is replaced by the Panet structure to improve the utilization of the underlying information; finally, Distance-IoU (DIoU) is used as the archor loss function to speed up the model convergence. The test results on the Flir image data set show that the improved model of the proposed improved YOLOv3 infrared recognition algorithm achieves better detection accuracy in terms of precision and recall when the model size is almost unchanged. Compared with YOLOv3, there are 2.94% and 3.12% increases in pedestrians and cars, respectively, and the average AP also increases by 3.03%. Experiments show that the improved method not only improves the detection accuracy, but also reduces the amount of models and improves the detection speed.

Keywords

Target Recognition, Pedestrian Vehicle Detection, YOLOv3, Backbone, Feature Fusion, Loss Function

1. 引言

目标识别检测技术是计算机视觉的重要分支，广泛应用于工业、军事，尤其是无人驾驶领域。在车辆、行人、交通标识等标志物的识别里发挥了重要作用。传统的目标检测算法大多使用可见光进行目标检测，在夜间、沙尘暴、雾雨天气等可见光资源较少、目标距离较远的情况下，很难进行有效的成像，检测效果欠佳，而红外影像技术基于检测物体的红外辐射能量进行成像，在无光的夜晚或者烟尘环境里依旧可以成像，受可见度低等因素影响较小，因而被广泛应用于夜间成像与非接触测温应用等多个领域。

传统的目标检测算法通常采用滑动窗口的方式，即采用一个窗口，在检测图片上滑动选取感兴趣区域，分别对滑动的每个窗口进行特征提取，如方向梯直方图(Histogram of Oriented Gradients, HOG) [1]、局部二值模式(Local Binary Pattern, LBP) [2]、尺度不变特征变换(Scale-Invariant Feature Transform, SIFT) [3]等，之后对提取的特征利用机器学习算法。但是由于不同图片的尺寸存在差异，如果使用固定窗口选取会存在重复、遗漏等问题，并且如果采用不同滑窗方式识别会出现计算量大，速度过慢的问题。

传统的红外影像目标检测方法一般有背景差分法、帧间差分法、光流法。Wren 等人[4]将图像中当前帧与已经确定或实时获取的背景图像作差，从而获得物体的位置大小等相关特征。Yin 等人[5]提出可以将相邻的几帧图像进行相减，并且对作差后的图像镜像阈值化来获得物体特征。Horn 和 Schunck 等人[6]利用图像上像素在时间上的变化及其相关性来计算物体运动状态。

近些年来，随着卷积网络的出现，自动提取特征，更高的准确率，更快的运算速度等特点使得基于深度学习的目标检测方法得到广泛研究。目前，基于深度学习的目标检测算法主要分为两类：二阶段(Two

stage)目标区域检测算法和一阶段(One stage)目标区域提取算法。基于区域检测的目标检测算法先进行区域生成，该区域称之为 region proposal (简称 RP，一个有可能包含待检物体的预选框) [7]，再通过卷积神经网络进行样本分类。常见 two stage 算法有：R-CNN [7]、SPP-Net [8]、Fast R-CNN [9]、Faster R-CNN [10]和 R-FCN [11]。基于区域提取的目标检测算法则是不用 RP，直接在网络中提取特征来预测物体分类和位置。常见的 one stage 目标检测算法有：OverFeat [12]、YOLOv1 [13]、YOLOv2 [14]、YOLOv3 [15]、SSD [16]和 RetinaNet [17]等。

YOLO 算法相较于 R-CNN 系列算法，仅通过一个卷积网络就能实现对目标物体的位置与类别的检测，凭借其实时性高、便于操作的特点得到了广泛应用。而 YOLOv3 作为 YOLO 系列的巅峰之作，不但延续了 YOLO 系列的便捷性，更是做出了改进。首先在特征提取部分采用 darknet-53 的 backbone 代替之前的 darknet-19，实现多尺度特征融合，在保证了实用性的同时，又提高了准确性。而在无人驾驶领域对于车辆和行人的目标检测，YOLOv3 的灵活性满足了笔者对于目标检测的需要，同时针对于目标检测算法在夜间环境下的检测精度较低的问题，采用 Flir 红外影像数据集，利用红外影像夜间成像的优势，对于 YOLOv3 的 backbone、特征融合与损失函数三方面进行微调改进，实现对于夜间车辆与行人的目标检测。

2. 改进的 YOLOv3 目标检测算法

2.1. YOLOv3 基本原理

YOLOv3 是一阶段(One stage)的目标检测算法，将输入图像划分成三种不同网格，分别对应目标检测的小尺寸目标，中尺寸目标，大尺寸目标。YOLOv3 的 backbone 采用 DarkNet-53 [15]的残差神经网络，能够在加深网络的同时又较好地解决梯度消失的问题，有助于数据训练和特征提取融合。YOLOv3 的网格结构如图 1 所示。在多尺度特征融合方面，YOLOv3 采用 FPN (feature pyramid networks) [18]特征融合结构。

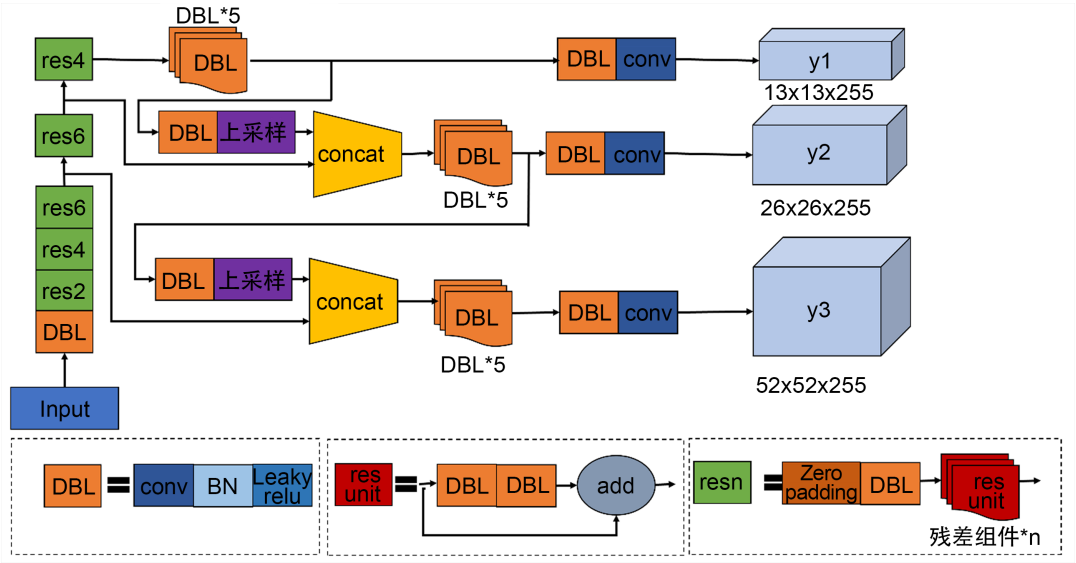

Figure 1. YOLOv3 structure diagram
图 1. YOLOv3 结构图

2.2. 深度可分离卷积

2017 年 Howard [20]等人提出了深度可分离卷积(depthwise separable convolutions)，通过分步卷积方

法大幅降低了模型的参数量。因此为了防止模型参数量过大的问题，本方法将后三个常规卷积更换为深度可分离卷积。深度可分离卷积主要分为两个过程，分别为逐通道卷积 DW (Depthwise Convolution)和逐点卷积 PW (Pointwise Convolution)。首先，图像输入经过第一次卷积运算，逐通道卷积的一个卷积核负责一个通道，卷积核的数量与上层的通道数一致，计算量为卷积核 W × 卷积核 H × (图片 W − 卷积核 W + 1) × (图片 H − 卷积核 H + 1) × 输入通道数。但逐通道卷积生成的 Feature map 数量与输入层的通道数相同，没有有效利用不同通道在相同空间位置上的 Feature 信息，因此需要后续逐点卷积对于生成的 Feature map 进行组合。逐点卷积部分则是进行 1 × 1 卷积，进行单点上的特征提取，将上一步生成的 Feature map 在深度方向上进行加权组合，生成新的 Feature map。其计算量为特征层 W×特征层 H×输入通道数 × 输出通道数。

将后续常规卷积更换为深度可分离卷积的方法，相较于常规卷积而言，在保证精度不会损失太多的情况下，大大降低了计算量与参数量。同时深度可分离卷积经常规卷积同时考虑区域和通道的拆分，深度可分离卷积先只考虑区域，然后再考虑通道。实现通道和区域的分离。

2.3. 模型改进

基于上述问题，文中针对于 YOLOv3 的 backbone、特征融合和损失函数三个部分进行优化改进，优化主干网络，增加大尺寸图像的卷积次数，提高特征提取能力。同时将部分普通卷积更换为深度可分离卷积，降低参数量。特征融合部分将 FPN 更换为 PANet [19]中的结构。在损失函数方面将 archor 中的定位误差改为 CIoU [21]，减少损失程度，增强信息丰富度。

2.3.1. Backbone

YOLOv3 的 backbone 采用的是由残差网络(Residual)组成的 Darknet-53，在加深网络的同时又能较好地解决梯度消失问题。本文为减少图像在卷积过程中损失过多信息，将前两个残差单元的卷积次数分别从 1，2 增至 2，4，提高大尺寸图像的特征提取能力，改进后的残差网络进行特征提取流程图如图 3。其中每个残差单元首尾连接，组成一个大残差边，这样较为有效地减少了传统卷积层在特征提取的过程中存在的信息丢失、损耗的问题，有效地保护了信息的完整性，以两个小残差单元为例，改进的残差单元结构如图 2 所示。

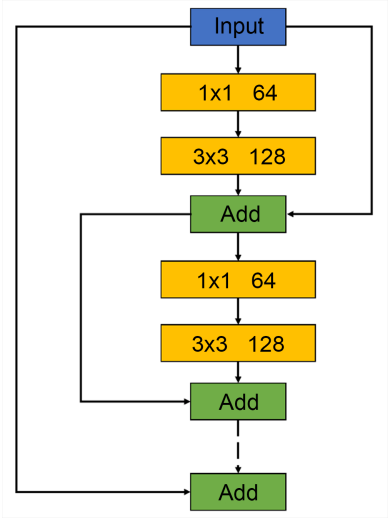

Figure 2. Improved residual unit structure diagram
图 2. 改进残差单元结构图

为了保证添加上述残差单元的模型参数量不至于过高，本文将后三个残差单元中的常规卷积更换为深度可分离卷积，较好地减少了参数量并提高了模型运行速度。改进后的残差网络特征提取流程如图 3。

	Type	Filters	Size		Output	
	Convolutional	32	3	3	256	256
	Convolutional	64	3		128	128
	Convolutional	32	1	1		
2	Convolutional	64	3	3		
	Residual				128	128
	Convolutional	128	3	3/2	64	64
	Convolutional	64	1	1		
4	Convolutional	128	3	3		
	Residual				64	64
	Convolutional	256	3	3/2	32	32
	Depthwise separable convolution	128	1	1		
6	Depthwise separable convolution	256	3	3		
	Residual				32	32
	Convolutional	512	3	3/2	16	16
	Depthwise separable convolution	256	1	1		
6	Depthwise separable convolution	512	3	3		
	Residual				16	16
	Convolutional	1024	3	3/2	8	8
	Depthwise separable convolution	512	1	1		
4	Depthwise separable convolution	1024	3	3		
	Residual				8	8
	Avgpool		Global			
	Connected		1000			
	Softmax					

Figure 3. Improved residual network feature extraction flowchart
图 3. 改进后残差网络特征提取流程

2.3.2. 特征融合

YOLOv3 的特征融合结构采用自下而上融合方式，由上层小尺寸特征图经上采样与下层大尺寸特征图融合，然后将融合后的三个尺寸特征图分别检测。而本文采用的是对比度较低的红外影像数据，相比于 RGB 格式的彩色图像，特征提取更加苦难。所以为了增加网络对提取特征的有效利用，将 YOLOv3 中的特征融合结构更换为 PANet 中的结构。PANet 结构在 FPN 的基础上又增加了一个从低层特征层到高层特征层的路径，将低层特征语义信息二次向上层传递，增加了底层特征语义信息的传递率与利用率，结构如图 4 所示。

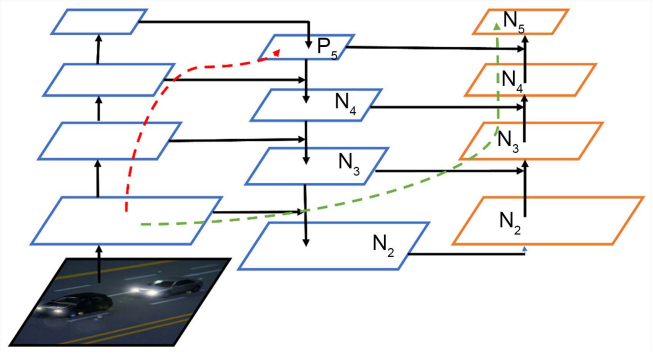

Figure 4. Improved feature fusion structure diagram
图 4. 改进特征融合结构图

2.3.3. 损失函数优化

YOLOv3 的损失函数由检测物体中心点的坐标误差(*x*, *y*)，anchor box 的宽高坐标误差(*w*, *h*)，置信度误差(confidence)，分类误差(class)组成。具体函数表示如下：

$$
\begin{aligned}
Loss = &-\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{x}_i^j \log\left(x_i^j\right)+\left(1-\hat{x}_i^j\right)\log\left(1-x_i^j\right)+\hat{y}_i^j \log\left(y_i^j\right)+\left(1-\hat{y}_i^j\right)\log\left(1-y_i^j\right)\right] \\
&+\lambda_{coord}\frac{1}{2}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\left(w_i^j-\hat{w}_i^j\right)^2+\left(h_i^j-\hat{h}_i^j\right)^2\right] \\
&-\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{C}_i^j \log\left(C_i^j\right)+\left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right] \\
&-\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{noobj}\left[\hat{C}_i^j \log\left(C_i^j\right)+\left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right] \\
&-\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\sum_{c\in classes}\left[\hat{P}_i^j \log\left(P_i^j\right)+\left(1-\hat{P}_i^j\right)\log\left(1-P_i^j\right)\right]
\end{aligned}
$$

$-\lambda_{coord}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{x}_i^j \log\left(x_i^j\right)+\left(1-\hat{x}_i^j\right)\log\left(1-x_i^j\right)+\hat{y}_i^j \log\left(y_i^j\right)+\left(1-\hat{y}_i^j\right)\log\left(1-y_i^j\right)\right]$为预测框中心坐标误差，其中$\left(x_i^j, y_i^j\right)$是 YOLOv3 的预测框中心点坐标，$\left(\hat{x}_i^j, \hat{y}_i^j\right)$为设定预测框的中心点坐标，$I_{ij}^{obj}$为预测框是否检测一个目标物体，其值为 1 或 0。

$\lambda_{coord}\frac{1}{2}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\left(w_i^j-\hat{w}_i^j\right)^2+\left(h_i^j-\hat{h}_i^j\right)^2\right]$为预测框宽高误差，其中$\left(w_i^j, h_i^j\right)$为预测框宽高，$\left(\hat{w}_i^j, \hat{h}_i^j\right)$为设定预测框的宽高。

$-\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{C}_i^j \log\left(C_i^j\right)+\left(1-\hat{C}_i^j\right)\log\left(1-C_i^j\right)\right]$为预测框置信度损失，$C_i^j$为预测框中是否含有目标检测物体的概率，$\hat{C}_i^j$为其真实值。

$-\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\sum_{c\in classes}\left[\hat{P}_i^j \log\left(P_i^j\right)+\left(1-\hat{P}_i^j\right)\log\left(1-P_i^j\right)\right]$为预测框中目标物体的类别损失，$P_i^j$表示预测框中目标检测物体是否属于检测类别的概率，$\hat{P}_i^j$其所属类别的真实值。

在实际训练中 anchor 坐标回归对模型收敛其到至关重要的作用，所以一个合理的损失函数可以加快模型收敛。所以本文将 YOLOv3 的 anchor 损失函数替换为 DIoU (Distance-IoU) [22]，DIoU 计算公式为：

$$
\mathcal{L}_{\mathrm{DIoU}} = 1-\mathrm{IoU}+\frac{\rho^2\left(b, b_{gt}\right)}{c^2}
$$

3. 实验结果分析与讨论

为了验证本文提出算法的有效性，选择由 Flir 公司发布的开源红外数据集进行对比实验。Flir 数据集于 2018 年 7 月发行，数据集包括同步注释热图像和无注释 RPG 图像供参考，共有 14,000 张图像，以 30 帧频率记录视频信息。注释标签包括行人、汽车、自行车、狗等标签。考虑到训练样本的平衡性，本文选取目标数量最多的汽车和行人构建训练数据集，共 7000 张。为了保证实验的公平性，本文采用完全相同的超参数对 YOLOv3 和改进模型分别进行训练，采用 AP 和 mAP 作为评价指标，对比结果如表 1 所示。

由表 1 可以看出本文改进算法在模型大小几乎不变的情况下，相比于 YOLOv3 在行人和汽车两类上分别有 2.94%和 3.12%提升，平均 AP 也有 3.03%的提升。为了更直观的展现改进算法识别能力，选取了

非训练集的红外影像 6 张进行实际识别，识别结果如图 5 所示。

Table 1. Comparative experimental results
表 1. 对比实验结果

models	Person (%)	Car (%)	mAP (%)	模型大小
YOLOv3	76.52	79.83	78.17	235M
本文模型	79.46	82.93	81.20	232M

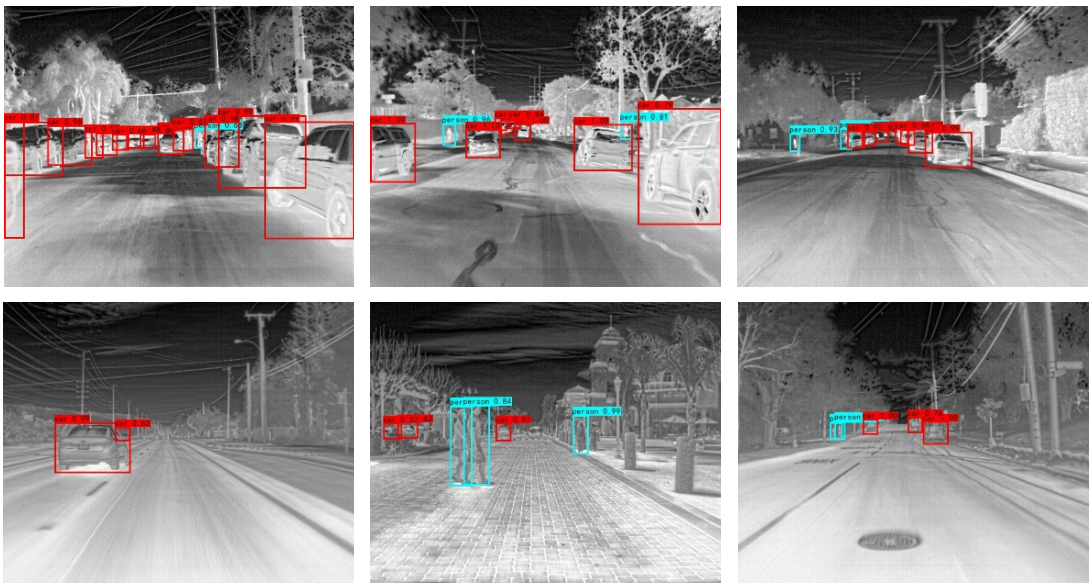

Figure 5. Recognition renderings
图 5. 识别效果图

从图 5 可以看出，改进的模型针对于不同道路情况，覆盖行人和机动车数量不同的各种情况，均实现了较好的识别效果，在识别率和召回率方面均有不错的表现，远处小目标没有出现漏检情况，人车密集的场所也能做到全部准确识别分类。

4. 结论

针对于夜间自动驾驶常规目标检测行人与车辆效果不佳的问题，本文提出一种基于 YOLOv3 算法的红外影像目标检测。该方法基于 YOLOv3 算法，优化调整其 backbone 结构，增加大尺寸图像的卷积次数，提高特征提取能力，并将后续常规卷积更换为深度可分离卷积，以减小模型参数量，加快模型运算速度。并将 YOLOv3 算法多尺度融合中的 FPN 模型更换为 PAnet 模型，加快底层语义信息的传递效率。同时在损失函数部分采用 DIoU 结构，加快模型收敛。并通过在 Flir 红外影像集上的测试，本文所提出的算法在模型量基本不变的情况下，相较于原始 YOLOv3 技术有较明显提升，对于行人和机动车检测率都有不错表现，平均 AP 提高了 3.06%。本文下一步工作将注重于夜间自动驾驶更多交通标志物的检测，同时进一步优化改进模型，提高目标检测效率的同时减少模型参数量。

参考文献

[1] Dalal, N. and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. *IEEE Computer Society Conference on Computer Vision & Pattern Recognition*, San Diego, 20-25 June 2005, 886-893.

[2] Ojala, T., Pietikäinen, M. and Harwood, D. (1996) A Comparative Study of Texture Measures with Classification Based on Feature Distributions. *Pattern Recognition*, **29**, 51-59. https://doi.org/10.1016/0031-3203(95)00067-4

[3] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[4] Wren, C.R., Azarbayejani, A.J., Darrell, T.J., *et al*. (1996) Pfinder: Real-Time Tracking of the Human Body. *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, Killington, 14-16 October 1996, 51-59.

[5] Yin, J., Lei, L., He, L., *et al*. (2016) The Infrared Moving Object Detection and Security Detection Related Algorithms Based on W4 and Frame Difference. *Infrared Physics & Technology*, **77**, 302-315. https://doi.org/10.1016/j.infrared.2016.06.004

[6] Horn, B.K.P. and Schunck, B.G. (1981) Determining Optical Flow. *Artificial Intelligence*, **17**, 185-203. https://doi.org/10.1016/0004-3702(81)90024-2

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

[8] He, K., Zhang, X., Ren, S., *et al*. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **37**, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

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

[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 and Machine Intelligence*, **39**, 1137-1149.

[11] Dai, J., Li, Y., He, K., *et al*. (2016) R-FCN: Object Detection via Region-Based Fully Convolutional Networks. *Proceedings of the* 30*th International Conference on Neural Information Processing Systems*, Barcelona, 379.

[12] Sermanet, P., Eigen, D., Zhang, X., *et al*. (2013) OverFeat: Integrated Recognition, Localization and Detection Using Convolutional Networks. arXiv:1312.6229.

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

[14] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[15] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement. arXiv: 1804.02767..

[16] Liu, W., Anguelov, D., Erhan, D., *et al*. (2016) SSD: Single Shot MultiBox Detector. Springer, Cham.

[17] Lin, T.Y., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 2999-3007. https://doi.org/10.1109/ICCV.2017.324

[18] Lin, T.Y., Dollar, P., Girshick, R., *et al*. (2017) Feature Pyramid Networks for Object Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 936-944. https://doi.org/10.1109/CVPR.2017.106

[19] Liu, S., Qi, L., Qin, H., *et al*. (2018) Path Aggregation Network for Instance Segmentation. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 8759-8768. https://doi.org/10.1109/CVPR.2018.00913

[20] Howard, A.G., Zhu, M., Chen, B., *et al*. (2017) MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861.

[21] Zheng, Z., Wang, P., Ren, D., *et al*. (2020) Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation. *IEEE Transactions on Cybernetics*. https://doi.org/10.1109/TCYB.2021.3095305

[22] Zheng, Z., Wang, P., Liu, W., *et al*. (2020) Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 12993-13000.