**Artificial Intelligence and Robotics Research** 人工智能与机器人研究, 2024, 13(1), 112-120
Published Online February 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.131013

Hans汉斯

# 基于特征重构的YOLO目标检测模型实现

**纪宇龙**

天津工业大学软件学院，天津

收稿日期：2023年11月29日；录用日期：2024年2月23日；发布日期：2024年2月29日

## 摘　要

**目标检测是计算机视觉领域的四大核心任务之一，它涵盖了目标的分类和定位，至今已有近二十年的研究历史。YOLOv5在YOLO系列目标检测算法中广泛应用，这得益于其稳定的工程实践能力，以及可以较好地平衡检测精度和速度，但YOLOv5算法的检测精度距两阶段目标检测器的性能还有差距。因此本文选取其作为研究对象，对其进行改进，力求在保证检测速度基本不变的同时尽可能地提升精度。针对YOLOv5s目标检测算法不能充分利用特征信息的问题，对YOLOv5s进行改进后，提出一种基于特征重构模块的目标检测模型(F-YOLOv5s)。该特征重构模块将w-h平面上的特征信息转移到空间维度，能够减少下采样带来的信息损失，从而提高目标检测的准确率。实验表明，在PASCAL VOC2007和VOC2012数据集上，本文提出的特征重构模块能有效提高特征信息的利用率，使得检测精度大幅度提升。**

## 关键词

**目标检测，YOLOv5s，特征重构**

# Implementation of YOLO Target Detection Model Based on Feature Reconstruction

**Yulong Ji**

College of Software, Tiangong University, Tianjin

Received: Nov. 29th, 2023; accepted: Feb. 23rd, 2024; published: Feb. 29th, 2024

## Abstract

**Object detection is one of the four core tasks in the field of computer vision, which covers the classification and localization of objects, and has been studied for nearly two decades. YOLOv5 is widely used in the YOLO series target detection algorithms, which is due to its stable engineering practice ability and good balance between detection accuracy and speed. However, the detection accuracy of YOLOv5 algorithm is still short of the performance of the two-stage target detector.**

文章引用: 纪宇龙. 基于特征重构的 YOLO 目标检测模型实现[J]. 人工智能与机器人研究, 2024, 13(1): 112-120.
DOI: 10.12677/airr.2024.131013

**Aiming at the problem that the YOLOv5 target detection algorithm cannot make full use of feature information, an improved YOLOv5 target detection model based on feature reconstruction module (F-YOLOv5) is proposed. The feature reconstruction module transfers feature information from the w-h plane to the spatial dimension, reducing information loss caused by down sampling and thereby improving the accuracy of object detection. Experiments show that on PASCAL VOC2007 and VOC2012 data sets, the feature reconstruction module proposed in this paper can effectively improve the utilization rate of feature information and greatly improve the detection accuracy.**

## Keywords

**Object Detection, YOLOv5s, Feature Reconstruction**



## 1. 引言

我们将可对图像中的目标进行识别和定位的技术称之为目标检测，这些目标包括人物、动物以及日常生活中常见的物品。目标检测技术应用广泛，在建筑工地用于监视以确保安全施工，在工业零件的生产中可用于检测出瑕疵产品，在医疗场景中可进行辅助诊断，在汽车行业可用于自动驾驶。在现实生活中，由于光照、遮挡所造成的阴影、成像品质等诸多因素，使目标检测具有更大的挑战性[1]。

目标检测算法大概发展为两个阶段，第一个阶段为传统的目标检测算法，人为进行提取特征和指定分类器来实现目标检测。特征提取主要采用尺度不变特征变换 SIFT [2]、局部二值模式 LBP [3]和方向梯度直方图 HOG [4]，分类器一般选择支持向量机 SVM [5]与 Adaboost [6]算法来实现任务分类功能。但是，传统的目标检测存在以下缺陷：(1) 虽然思路简洁，但实则计算开销巨大；(2) 图像语义信息非常复杂，人工提取难以获得，导致特征表达能力不足，不同任务之间鲁棒性较差[7]。

第二个阶段为基于深度学习的目标检测算法。随着卷积神经网络快速发展，人们提出用卷积神经网络提取图像特征[8]，对比传统的目标检测算法，基于深度学习的目标检测算法不仅漏检率和错误率低，检测精度还更高。现在，基于深度学习的目标检测算法主要分为两大类别。一类是双阶段目标检测，将候选框提取和分类顺序进行，先使用 RPN [9]提取目标的候选框，再对区域位置校准后进行分类得到最终的检测结果。由于这种目标检测网络分两步执行，注重高准确率，复杂的结构导致检测速度较慢，其中典型代表有第一次将候选区与卷积神经网络相结合的 R-CNN [10]、基于 R-CNN 改进，添加金字塔池化层的 SPPNet [11]等；另一类是单阶段目标检测，其抛弃了生成候选区域的阶段，将目标检测当作回归问题来解决，经过单次检测即可产生类别概率和位置坐标。由于这种目标检测网络在结构设计上进行简化，更加注重实时性，导致检测精度较低[12]，其中典型代表有 YOLO [13] [14] [15] [16]和 Retina-Net [17]等。

本文针对在 YOLOv5s 的基础上进行了研究和改进，提出了特征重构模块，减少特征损失，提高特征信息的利用率，并将特征重构模块移入 YOLOv5s 网络中，使用特征模块，保留更多的特征信息，提高了检测的准确率。

## 2. 相关工作

古人云：“知秋一叶，尝鼎一脔”，其中就蕴含着采样的思想。采样，顾名思义，就是从特定的概率分布中抽取相应样本点的过程。而在深度学习中，它可以将复杂的分布简化为离散的样本点，用于对

样本集进行调整以更好的适应后期的模型学习。采样又可以分成下采样和上采样。

在卷积神经网络中，由于输入图像通过卷积神经网络(CNN)提取特征后，输出的尺寸往往会变小，而有时我们需要将图像恢复到原来的尺寸以便进行进一步的计算(如图像的语义分割)，这个使图像由小分辨率映射到大分辨率的操作，叫做上采样。下采样就是指的是成比例缩小特征图宽和高的过程，比如从(W, H)变为(W/2, H/2)，即对卷积得到的 Feature Map 进行进一步压缩。下采样的作用有很多，包括：降低维度、减少网络要学习的参数数量、防止过拟合、增大感受野，使得后面的卷积核能够学到更加全局的信息。

池化操作是最早接触到的下采样方式。其中平均池化和最大池化两种池化方式最为常见。平均池化有种平滑滤波的味道，通过求取滑窗内的元素平均值作为当前特征点，根据滑窗的尺寸控制下采样的力度，尺寸越大采样率越高，但是边缘信息损失越大。最大池化类似锐化滤波，突出滑窗内的细节点。但是不论哪种池化操作，都是以牺牲部分信息为代价，换取数据量的减少。

步长大于 1 的卷积也可以实现下采样的作用。卷积操作可以获得图像像素之间的特征相关性，采用步长大于 1 的跳跃可以实现数据降维，但是跳跃采样造成的相邻像素点特征丢失可能影响最终效果。卷积实现的下采样和池化相比，池化操作提供了一种非线性，这种非线性需要较深的卷积叠加才能实现，因此当网络比较浅的时候，池化有一定优势；但是当网络很深的时候，多层叠加的卷积可以学到池化所能提供的非线性，甚至能根据训练集学到比池化更好的非线性，因此当网络比较深的时候，不使用池化没多大关系，甚至更好。另外，池化下采样比较粗暴，可能将有用的信息滤除掉，而卷积下采样过程控制了步进大小，信息融合比较好，现在池化操作大部分被卷积所替代。

在老版本的 YOLOv5 中，也出现过一种结构叫做 Focus。Focus 模块采用切片操作把高分辨率的图片(特征图)拆分成多个低分辨率的图片(特征图)，即隔列采样 + 拼接，再经过一次 1 × 1 卷积操作，用来改变通道数，最终图片尺寸下降一半，通道数变为卷积后的输出通道数。原作者设计 Focus 的目的是减少参数量，并增加计算速度，并不会提升目标检测模型精度。

## 3. 模型

### 3.1. YOLOv5s 网络

本文在 YOLOv5s 网络基础上改进。YOLOv5s 的网络结构分为 Backbone、Neck、Head，如图 1 所示。

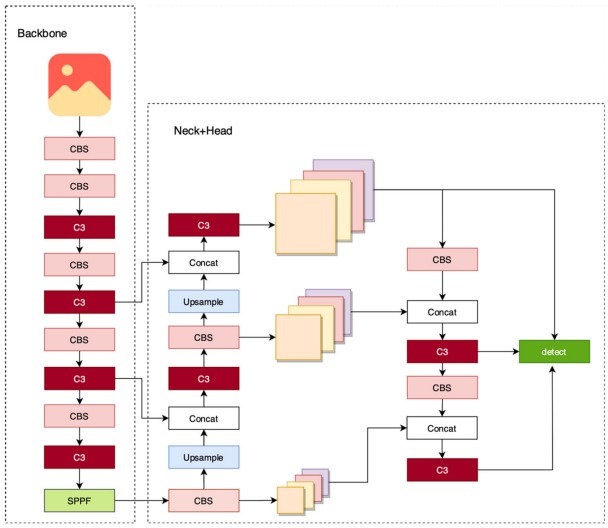


**Figure 1.** YOLOv5s network architecture
**图 1.** YOLOv5s 网络结构

Backbone 由 CBS、C3 以及 SPPF 构成，通过五次 CBS 提取输入的图像特征，每次卷积后利用 C3 模块进行特征堆叠加深网络，并在 Backbone 末尾添加 SPPF 模块进行池化特征。

CBS 卷积模块由 Conv 卷积、Batch Norm 标准化和 Silu 激活函数组成，主要作用是增加模型的非线性拟合能力。

C3 模块一共包含三次 CBS 卷积与若干次 Bottle Neck，借鉴 ResNet 残差网络思想，将输入特征层拆分为两部分处理。主干部分利用 CBS 与 Bottle Neck 逐步提取层内特征，加深网络并增加网络的感受野，分支部分仅用单次 CBS 调整空间分辨率，最后再由第三个 CBS 提取 Concat 后的新特征层。

在 SPPF 模块中，特征图串行通过一个 CBS 和三个 5x5 池化层，用 Concat 模块堆叠前面四个结果后再经过一个 CBS 层处理，将同一特征图不同尺度下的特征融合到一起，丰富特征图的语义特征。

Neck 部分采用了 FPN [18]与 PAN [19]结构来进行融合，PAN 结构中采用自底向上和从上到下的两条路径上的特征融合。自底向上的路径为通过 CBS 卷积和上采样，由 20 × 20 的深层次特征通过 Concat 进行堆叠至 40 × 40、80 × 80 的中低层次特征。从上到下的路径则相反，将低层次特征通过 CBS 卷积进行提取并下采样到中高层次特征，再利用 C3 层将 Concat 堆叠后的特征进行特征融合，使深层次的语义特征与浅层次的位置信息特征相互补充，提高模型的特征表达能力。

Head 部分根据 Neck 的三种不同尺度的特征图将图片划分成不同的网格尺度，在每个网格中设置不同宽高比和不同大小的三个 Anchor (先验框)来检测目标物，一共聚类得到 9 个不同尺寸的先验框。由于分辨率最低的特征图由深层网络卷积得来，局部感受野最大，适合大物体检测，分辨率中等的特征图适合检测中等大小物体，相反的，分辨率最高的特征图是由浅层网络卷积得来，其感受野最小，因而比较适合小物体检测，最后根据位置信息调整 Anchor 宽高比，生成真实检测框。

### 3.2. F-YOLOv5s 网络

YOLOv5s 目标检测模型遵循近几年下采样的方式，采用了步长大于 1 的卷积的形式进行下采样。Backbone 中有 5 个单独的 CBS 模块，代表下采样了五次，设计在 3 个不同尺度的特征图上来进行物体检测，取下采样倍数分别为 32 倍、16 倍和 8 倍的输出张量输入到 FPN 结构中，采取上采样的形式进行特征融合，之后再进入到 PAN 结构，采取下采样的形式进行增强特征融合。

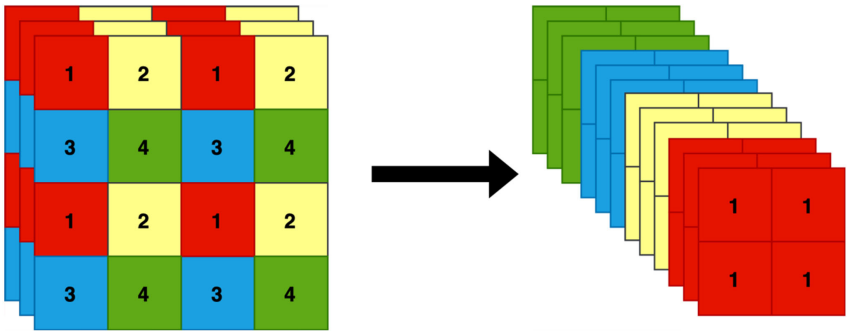


**Figure 2.** Feature reconstruction module
**图 2.** 特征重构模块

虽然步长大于 1 的卷积操作比池化操作的带来的信息损失要小，但面对大物体时，感受野最大的依然可能会丢失高级语义信息，面对小物体时，下采样最小的输出尺度也可能丧失强定位特征。因此，为了获取更加全面的更加丰富的特征信息，本文借鉴 Focus 的思想提出了特征重构(Feature_Reconstruction)模块，它是一种特殊的下采样方式，如图 2 所示，将 4 × 4 × 3 的 Tensor 通过间隔采样拆分成 4 分，在通

道维度上拼接生成 2 × 2 × 2 的 Tensor，这样宽高减半，通道为原来的四倍，采用这种方式可以减少下采样带来的信息损失，更详细得来说，是对图片进行切片操作，具体操作是在一张图片中每隔一个像素拿到一个值，类似于邻近下采样，这样就拿到了四张图片，四张图片互补，但是没有信息丢失，这样一来，将 W、H 信息就集中到了通道空间，输入通道扩充了 4 倍，即拼接起来的图片相对于原先的 RGB 三通道模式变成了 12 个通道，这样再对其后续操作，得到没有信息丢失后的新图片。

结合特征重构模块，对 YOLOv5s 网络模型的改进思路有主要有两方面，改进后的网络结构如图 3 所示。

一方面是对 Backbone 进行改进。首先把第一层的 6 × 6 的 CBS 模块换成 Focus 模块，在理论上来说，6 × 6 的 CBS 模块和 Focus 模块是等价的，它们的计算量和参数量是相等的，其实，这里的替换是和特征重构模块有关，经过实验比较，Focus 模块和特征重构模块一起使用，能达到更好的性能。其次，在每个 C3 模块前面添加特征重构模块，C3 模块是 Backbone 对残差特征进行学习的主要模块，在每个 C3 模块前面添加特征重构模块，能使输入 C3 模块的特征信息更加丰富，这样 C3 学到的特征信息也会更多，不管是低层次语义特征，还是高层次定位特征都会有所增加，能全方面提升目标检测模型的精度。除此之外，还修改 Backbone 中所有 CBS 模块的卷积的步长为 1，使得 CBS 模块只用来进行提取特征，把下采样的任务交给特征重构模块进行。

另一方面是对 PAN 结构的改进。为了能更加充分的利用浅层特征图的有效信息，本文在 PAN 结构中的每个 Concat 模块前面加入特征重构模块，在新的特征融合结构下，待检测特征图包含更多的几何信息，这些低级的几何信息可以帮助物体更准确地定位。

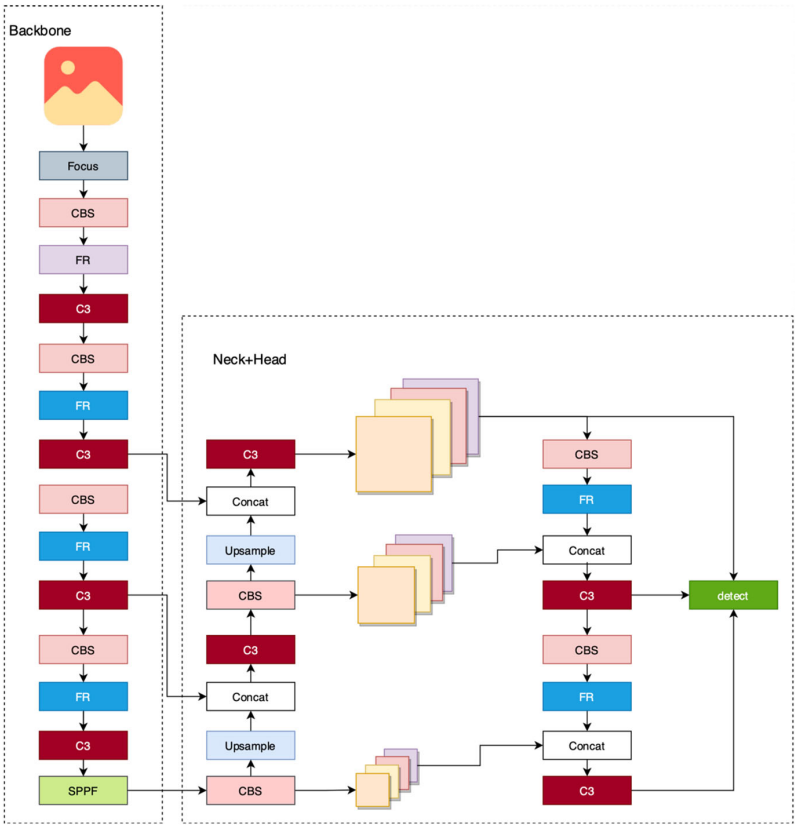


**Figure 3.** F-YOLOv5s network architecture
**图 3.** F-YOLOv5s 网络结构

# 4. 实验

## 4.1. 数据集和参数设置

为了验证本文提出的 F-YOLOv5s 的有效性，进行实验，使用 PASCAL VOC 数据集进行训练和测试。在本实验中，图片采用 640 × 640 作为网络输入，训练集选用 PASCAL VOC 2007 和 PASCAL VOC 2012 数据集中的 trainval 部分，共 16551 张图片；测试集选用 VOC 2007 数据集中的 test 部分，共 4952 张图片。

本文的实验是在一台 Ubuntu16.04 系统的 GPU 服务器上进行，显卡为 RTX A5000，开发语言是 Python，框架是 PyTorch，续联采用了 Amd 优化器进行参数优化。训练网络时，超参数的设置为选择 hyp.scratch-high.yaml 文件，batch_size 为 32，迭代总批次为 300，学习率采用余弦退火衰减来保证模型更好的收敛。

## 4.2. 改进前后结果对比

将训练后的网络在 PASCAL VOC 2007 测试集上进行测试，绘制了召回率－精确度曲线图，如图 4、图 5 所示。横坐标 Recall 表示召回率，纵坐标 Precision 表示精度，精确率与召回率这两个评估指标通常是相互矛盾的，通常情况下，会使用 PR 曲线来表示分类器在精确率与召回率之间的平衡。改进后的模型对各个类别的精度均有提升，并且数据集中所有类别的平均准确率达到了 78.5%。

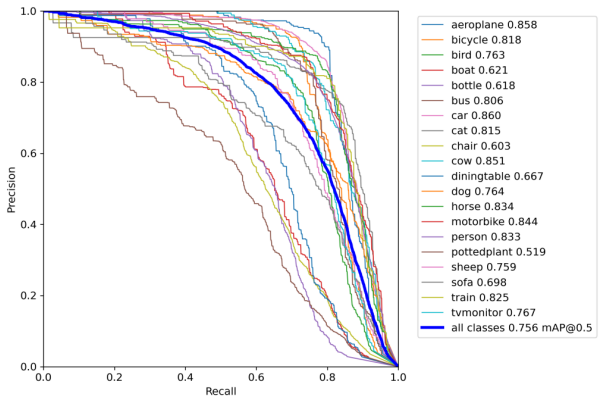


**Figure 4.** YOLOv5s
**图 4.** YOLOv5s

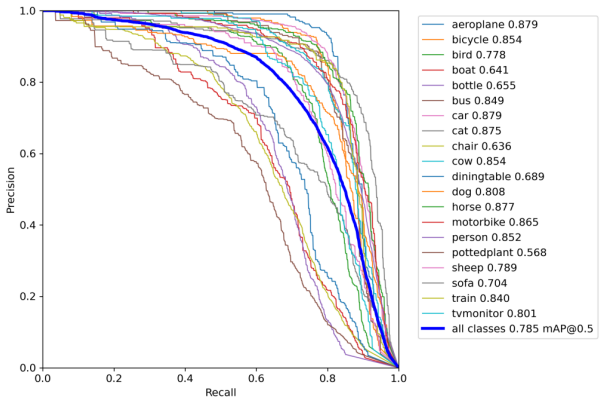


**Figure 5.** F-YOLOv5s
**图 5.** F-YOLOv5s

为了证明所提出模块的有效性，分别对 YOLOv5s、YOLOv5s + Focus、F-YOLOv5s 三种模型在数据集上的 map 以及其他性能指标进行了测试。如表 1 所示，其中最优值已被加粗显示。

**Table 1.** Performance comparison table
**表 1.** 性能对比表

| | YOLOv5s (%) | YOLOv5s + Focus (%) | F-YOLOv5s (%) |
|---|---|---|---|
| mAP@.5 | 75.6 | 75.4 | **78.5** |
| mAP@0.5:0.95 | 49.2 | 49.3 | **52.4** |
| Precision | 78.2 | 78.4 | **80.7** |
| Recall | 68.3 | 67.9 | **71.5** |
| F1-score | 73.5 | 73.1 | **76.1** |

实验结果表明，本文提出的 F-YOLOv5s 模型相较于其他两种模型在性能指标方面均有所提升。当 IOU 阈值为 0.5 时，本文方法相较于原始 YOLOv5s 模型的 map 上升了 2.9%，相较于 YOLOv5s + Focus 模型的 map 上升了 3.1%。当 IOU 阈值在区间[0.5:0.95]时，本文提出模型比另外两种模型分别提升了 3.2%、3.1%。在检测精度方面，本文提出模型比另外两种模型分别提升了 2.5%、2.3%。在召回率方面，也是有较大提升，分别为 2.6%、3%。

### 4.3. 与其他算法结果对比

为了体现出本文算法的优越性，本文将改进后的网络与近年来其他目标检测网络进行对比，结果如表 2 所示，其中最优值已被加粗。

其中，以 ResNet-152 为 Backbone 的 PS-DK 网络，鉴于使用了相当大且深层的 Backbone 网络，使得自身的检测准确率到了最高，为 79.5%，本文提出的网络的准确率比其低了 1%，但网络的参数量更少，仅仅大约为 PS-DK 网络参数量的 1/9。对于 YOLOv7tiny + CIOU 来说，本文所提出网络的参数量虽然有所增加，但检测准确率比其提高了 14.06%。综合结果表明，改进后的目标检测网络与近年来其他的先进目标检测网络对比，也能展现出不错的性能。

**Table 2.** Comparison table of different object detection algorithms
**表 2.** 不同目标检测算法的对比表

| | Framework | Params (M) | Map@.5 |
|---|---|---|---|
| Faster R-CNN [14] | VGG-16 | 138 | 73.2 |
| PS-DK [20] | ResNet-152 | 215 | **79.5** |
| FCOS (Mutual Guidance) [21] | VGG-16 | 142 | 79.4 |
| YOLOv7tiny + CIOU [22] | YOLOv7tiny | **6.2** | 64.44 |
| F-YOLOv5s | YOLOv5s | 25 | 78.5 |

## 5. 结束语

本文研究基于 YOLOv5s，提出来一种改进的目标检测算法 F-YOLOv5s，提高检测网络的准确度。提出特征重构模块进行来进行下采样操作，使得特征信息不会丢失，提高特征信息的利用率。YOLOv5s 目标检测网络的骨干网络和脖颈部分引入特征重构模块，大幅度提高网络的准确率。改进后的网络在 PASCAL VOC 数据集上测试，较原始网络其准确率提升了 2.9%，达到了 78.5%，以及其他性能指标也均有所上升。接下来将继续优化改进该网络，研究注意力机制对该网络的影响。

## 参考文献


[1] 谷永立, 宗欣欣. 基于深度学习的目标检测研究综述[J]. 现代信息科技, 2022, 6(11): 76-81.

[2] Gao, X., Wu, Y., Yang, K., *et al*. (2015) Vehicle Bottom Anomaly Detection Algorithm Based on SIFT. *Optik*, **126**, 3562-3566. https://doi.org/10.1016/j.ijleo.2015.08.268

[3] 裘莉娅, 陈玮琳, 李范鸣, 等. 复杂背景下基于 LBP 纹理特征的运动目标快速检测算法[J]. 红外与毫米波学报, 2023, 41(3): 639-651.

[4] Chacon-Murguia, M.I., Rivero-Olivas, A. and Ramirez-Quintana, J.A. (2021) Adaptive Fuzzy Weighted Color Histogram and HOG Appearance Model for Object Tracking with a Dynamic Trained Neural Network Prediction. *Signal, Image and Video Processing*, **15**, 1585-1592. https://doi.org/10.1007/s11760-021-01891-9

[5] 李雄飞, 王婧, 张小利, 等. 基于 SVM 和窗口梯度的多焦距图像融合方法[J]. 吉林大学学报(工学版), 2020, 50(1): 227-236.

[6] Mehmood, Z. and Asghar, S. (2021) Customizing SVM as a Base Learner with AdaBoost Ensemble to Learn from Multi-Class Problems: A Hybrid Approach AdaBoost-MSVM. *Knowledge-Based Systems*, **217**, Article ID: 106845. https://doi.org/10.1016/j.knosys.2021.106845

[7] 刘国特, 伍伟权, 郭芳, 等. 基于改进级联 Gentle Adaboost 分类器的支柱绝缘子红外图像 AI 识别[J]. 高电压技术, 2022, 48(3): 1088-1095.

[8] He, X., Yang, H., Hu, Z., *et al*. (2022) Robust Lane Change Decision Making for Autonomous Vehicles: An Observation Adversarial Reinforcement Learning Approach. *IEEE Transactions on Intelligent Vehicles*, **8**, 184-193. https://doi.org/10.1109/TIV.2022.3165178

[9] Cho, M.A., Chung, T., Lee, H., *et al*. (2019) N-RPN: Hard Example Learning for Region Proposal Networks. 2019 *IEEE International Conference on Image Processing* (*ICIP*), Taipei, 22-25 September 2019, 3955-3959. https://doi.org/10.1109/ICIP.2019.8803519

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

[11] He, K., Zhang, X., Ren, S., *et al*. (2015) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[12] He, X., Lou, B., Yang, H., *et al*. (2022) Robust Decision Making for Autonomous Vehicles at Highway On-Ramps: A Constrained Adversarial Reinforcement Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, **24**, 4103-4113. https://doi.org/10.1109/TITS.2022.3229518

[13] Redmon, J., Divvala, S., Girshick, R., *et al*. (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

[14] Redmon, J. and Farhadi, A. (2017) YOLO9000: Better, Faster, Stronger. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 7263-7271. https://doi.org/10.1109/CVPR.2017.690

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

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

[17] Lin, T., Goyal, P., Girshick, R., *et al*. (2017) Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2980-2988. https://doi.org/10.1109/ICCV.2017.324

[18] Lin, T.Y., Dollár, P., Girshick, R., *et al*. (2017) Feature Pyramid Networks for Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 2117-2125. https://doi.org/10.1109/CVPR.2017.106

[19] Liu, S., Qi, L., Qin, H., *et al*. (2018) Path Aggregation Network for Instance Segmentation. *Proceedings of the IEEE 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] Kim, K., Ji, B.M., Yoon, D., *et al*. (2021) Self-Knowledge Distillation with Progressive Refinement of Targets. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 6567-6576. https://doi.org/10.1109/ICCV48922.2021.00650

[21] Zhang, H., Fromont, E., Lefèvre, S., *et al.* (2020) Localize to Classify and Classify to Localize: Mutual Guidance in

Object Detection. *Proceedings of the Asian Conference on Computer Vision*, Kyoto, 30 November-4 December 2020, 104-118.

[22] Zhang, H., Xu, C. and Zhang, S. (2023) Inner-IoU: More Effective Intersection over Union Loss with Auxiliary Bounding Box. arXiv: 2311.02877.