

引用格式:程钰清,贺占庄,马钟,等. 面向嵌入式 FPGA 的智能目标检测算法[J]. 微电子学与计算机, 2021, 38(6): 87-92.
[CHENG Y Q, HE Z Z, MA Z, et al. Intelligent target detection algorithm for embedded FPGA[J]. Microelectronics & Computer, 2021, 38(6): 87-92.]

面向嵌入式 FPGA 的智能目标检测算法

程钰清, 贺占庄, 马 钟, 毕瑞星, 毛远宏
(西安微电子技术研究所, 陕西 西安 710054)

摘 要: 随着识别率和实时性的提高, 卷积神经网络目标检测算法的计算复杂度和内存需求急剧增加, 难以应用在小尺寸和低功耗的嵌入式平台上. 本文在分析现有目标检测神经网络模型结构的基础上, 根据 FPGA 高实时性、低功耗以及并行处理的特点, 提出了一种在 FPGA 上高速运算的神经网络模型规整化方法. 在此方法指导下设计改进了一款目标检测神经网络模型结构, 包括删除 LRN 层、Scale 层的融合和替换 Leaky-ReLU 为 ReLU. 通过在 voc2007 数据集上的对比实验验证了算法结构的有效性, 在 PC 上其速度相比传统 YOLO-V1 算法提升了 11.5%. 在 Xilinx ZCU102 开发板上的仿真表明: 该改进的目标检测算法速度达到 29 FPS(Frames Per Second), 精度达到 62.3 mAP.

关键词: FPGA; 目标检测; YOLO 算法; 卷积神经网络
中图分类号: TP391.41; TP183 **文献标识码:** A **文章编号:** 1000-7180(2021)06-0087-06

Intelligent target detection algorithm for embedded FPGA

CHENG Yu-qing, HE Zhan-zhuang, MA Zhong, BI Rui-xing, MAO Yuan-hong
(Xi'an Microelectronics Technology Institute, Xi'an 710054, Shaanxi China)

Abstract: With the improvement of recognition rate and real-time performance, the computational complexity and memory requirements of convolutional neural network target detection algorithm increase sharply, which makes it difficult to be applied to embedded platform with small size and low power consumption. In this paper, based on the analysis of the existing neural network model structure of target detection, according to the characteristics of high real-time performance, low power consumption and parallel processing of FPGA, a neural network model normalization method based on high speed operation on FPGA is proposed. Under the guidance of this method, a target detection neural network model structure is designed and implemented, including removing the LRN layer, fusion of Scale layer and replacing Leaky-ReLU with ReLU. The effectiveness of the proposed algorithm structure is verified by comparative experiments on VOC2007 dataset. Compared with traditional YOLO-V1 algorithm, the speed of the proposed algorithm on PC is improved by 11.5%. Hardware simulation under Xilinx ZCU102 shows that the improved target detection algorithm can reach the speed of 29FPS (Frames Per Second) and the accuracy of 62.3mAP.

Key words: FPGA; object detection; YOLO algorithm; convolution neural network

1 引言

目标检测算法是计算机视觉领域的一个重要研

究方向^[1-3], 近年来, 基于卷积神经网络的目标检测算法取得了巨大突破, 在自动驾驶、人脸识别^[4]、行人检测等领域都获得了广泛的应用. 然而传统的卷

积神经网络目标检测算法^[5]计算冗杂、网络参数繁多,使得其只能在高性能计算机上运行^[6],从而难以满足实际应用需求。主流的解决方法之一是采用 FPGA 提高运算速度。FPGA 高实时性、低功耗以及并行处理等特点,使其能够完成多种情形的工作,具有较好的实用性和灵活性,适用于智能目标检测算法的各种应用场景。

但直接将卷积神经网络目标检测算法移植到 FPGA 上会造成运算速度的急剧下降,达不到实时性检测要求,这是因为卷积神经网络目标检测算法的网络结构种类较多,不能充分地利用 FPGA 的资源。针对这一问题,本文提出了一种在 FPGA 上高速运算的神经网络模型规整化方法^[7-8],并在此方法指导下设计改进了一款目标检测神经网络模型结构。该模型结构以 YOLO 算法模型^[9]为基础,采用 Inception 结构,并对模型结构进行部分适应性修改,包括删除 LRN 层、Scale 层的融合以及 Leaky-ReLU 层替换为 ReLU 层。在 PC 端,该模型结构在 voc2007 数据集上速度较传统 YOLO 算法提升了 11.5%。在 Xilinx ZCU102 开发平台上对训练后的模型进行部署,通过规整化计算方式,在功耗仅为 30w 下,达到了 62.3 mAP 的精确度和 29 FPS 的计算速度。

2 FPGA 上高速运算的神经网络模型规整化方法

由于直接将神经网络目标检测算法移植到 FPGA 会造成资源的大量浪费,无法满足实际应用需求。故为了在 FPGA 平台上利用其有限的计算资源实现可接受的目标检测算法实时性能,本文提出了一种在 FPGA 上高速运算的神经网络模型规整化方法,该方法主要对 CNN 模型结构有以下几点要求。

(1) 规整性

FPGA 擅长处理高度并行计算。在数据流方式下,FPGA 往往被用于进行分布式和并行计算,以达到充分利用有限的硬件资源、加速运算的目的。因此,非常规整的神经网络模型结构是进行模型规整化的首选。其中,存在问题的模型结构包括单独层(需要额外硬件资源),LRN 层(不同的输出映射相互影响),卷积层有大量内核大小不同的层(可能需要通过不同的方式加速)和过于复杂的网络结构。

(2) 全卷积网络结构

一个只包含卷积层并不包含任何全连接层的网络称为全卷积。全卷积网络使用更少的内存带宽。虽

然卷积层中的权重被重复使用多次,但全连接层需要为每个单独的多累积操作加载一个新的权重。由于在通用 FPGA 中内存带宽有限,所以在规整化神经网络结构设计中,需要考虑更高的计算-通信比率。

(3) 低计算复杂度

实时性是智能目标检测需求之一,虽然在检测时没有固定的帧率设定,但在实际应用程序中,大约 10FPS 就能满足实时需求。这使得移植到 FPGA 上,对目标检测算法进行运算时,在每一个前推中都应该有超过 230 亿的 MACCs,从而要求 CNN 智能目标检测模型结构具有较低的计算复杂度。

(4) 运算速度

尽管移植到 FPGA 上有许多限制,智能目标检测算法仍然应该提供高质量的检测结果,并应该在精度几乎不损失的前提下,对其检测速度进行优化。

3 基于规整化方法的神经网络模型结构设计

3.1 改进目标检测算法模型的选择

现有的基于卷积神经网络的目标检测算法主要分为两大类:基于候选区域的方法和基于回归的方法。但总体上,基于候选区域的目标检测算法实时性较差,很难满足实际需求。而基于回归的目标检测算法,由于将特征提取,目标分类和目标回归的所有过程都在一个卷积神经网络中完成,极大的改善了目标检测算法的实时性,故其使目标检测算法在嵌入式平台上的应用成为可能。

基于回归的目标检测算法中,性能较好的有 YOLO 算法系列,该系列算法已经发展到 V3 版本,但在 V2 和 V3 版本中^[12-13],存在自定义层,且自定义层在网络中间,FPGA 推断中的自定义层通过 CPU 运算实现,速度低,不符合本文提出的规整化方法。同时,在神经网络计算中,FPGA 与 CPU 频繁大量的数据交换,容易造成数据阻塞,降低了算法的运行速度。故本文最终选择 YOLO-V1 算法作为基础进行模型结构改进。但是传统 YOLO-V1 模型结构中存在 2 个全连接层并且是串行设计,显然这并不符合提出的规整化方法中全卷积网络结构要求,故本文最终选择改进的模型结构为 GoogLeNet-YOLO-V1。

3.2 高速度目标检测神经网络模型结构

从本文所提出的规整化方法来看,由于深度卷积计算多样性的本质,图 1(a)显示的 YOLO-V1 算法神经网络结构将无法充分利用 FPGA 的计算资

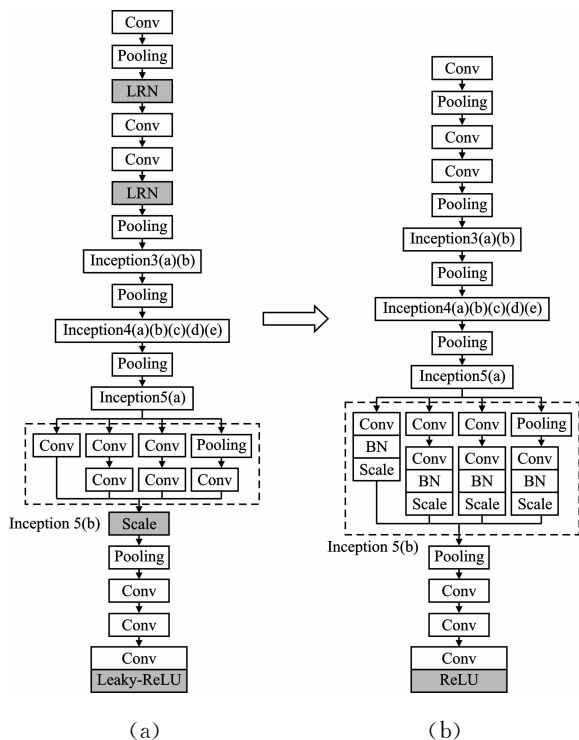


图1 YOLOv1 模型结构和改进的算法模型结构

(a)表示 YOLOv1 模型结构,其中最后一个 Conv 层激活函数是 Leaky-ReLU,其余 Conv 层均为 ReLU 激活函数(b)表示改进的算法模型结构,所有 Conv 层均使用 ReLU 激活函数

源,并且还需分配单独的资源实现 LRN 层、Scale 层以及 Leaky-ReLU 激活函数的计算。

因此,在图 1(b)中,有三次主要变化使得本文改进的目标检测算法模型结构适用于移植到 FPGA 上,在几乎不损失精度的条件下,计算速度达到了 29 FPS。

3.2.1 Scale 层的融合

如图 1(a)所示, YOLO-V1 算法神经网络结构中只使用了 1 个 Scale 层,不符合规整化方法中提到的规整性要求,影响计算效率。考虑 Scale 层在原始模型结构中的功能,本文将其加入到 Inception5(b)的各个尺度的卷积层中,并在网络推断时,参考 Caffe 中的 Batch Normalization 结构设计,将 Scale 层和 BN 层融合到卷积层中,替换原来卷积层中的 weight 和 bias 参数。

该融合策略的初衷是在不改变算法精度的前提下,提升运算速度。与 YOLO-V1 输出数据分布相比,本文提出的 Scale 层融合策略对输出数据分布的推导如下。

如图 2 所示,对于 YOLO-V1 算法神经网络结构最后一个 Inception 结构,以单独卷积 Conv1x1 为例,假设输出数据满足 $x \sim N(\mu, \sigma^2)$ 。

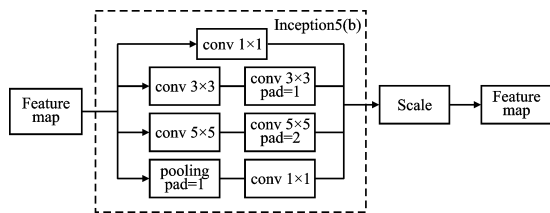


图2 YOLOv1 模型 Inception 结构

经过 Scale 层对数据做线性变化,进行比例缩放和位移,其公式如下:

$$y = \gamma x + \beta \quad (1)$$

式中, γ, β 表示自定义参数。需要注意的是,这两个参数均为向量,是针对输入的通道进行处理,并且通过迭代学习更新参数值。

最终输出数据满足 $X \sim N(\gamma\mu + \beta, (\gamma\sigma)^2)$ 。在 YOLO-V1 中,设定 $\gamma = 0.1, \beta = 0$,故数据分布满足 $X \sim N(0.1\mu, (0.1)^2\sigma^2)$

而如图 1(b)所示的本文提出的算法网络框架,单独卷积 Conv1x1 经过 BN 层和 Scale 层的输出推理如下。其中, BN 层为对数据分布进行归一化处理,其公式为

$$(x - \mu) / \sqrt{\sigma^2 + \epsilon} \quad (2)$$

式中, μ 表示输入数据的均值; σ^2 表示输入数据的方差; ϵ 表示滑动系数。经过该层计算,输出数据分布近似满足 $x \sim N(0, 1)$ 。再经过 Scale 层计算后,将卷积层输出值 $y = w * x + b$ 带入公式(1)和(2)中,合并后的结果如下

$$w_{\text{new}} = w_{\text{old}} * \gamma / \sqrt{\sigma^2 + \epsilon} \quad (3)$$

$$b_{\text{new}} = b_{\text{old}} \gamma / \sqrt{\sigma^2 + \epsilon} - \mu \gamma / \sqrt{\sigma^2 + \epsilon} + \beta \quad (4)$$

$$y = w_{\text{new}} * x + b_{\text{new}} \quad (5)$$

其输出数据分布满足 $x \sim N(\beta, \gamma^2)$, 本文设定的 γ, β 数值与 YOLO-V1 相同,故数据分布最终满足 $x \sim N(0, (0.1)^2)$ 。

与图 2 中 YOLO-V1 算法神经网络结构输出结果相比,本文改进的模型结构输出只是对数据进行归一化处理,并没有影响算法精度,但计算复杂度直线下降,运算速度大幅度提升,并且满足规整化方法,故将原始 Scale 层进行融合。

3.2.2 删除 LRN 层

如图 1(a)所示, YOLO-V1 算法神经网络结构中只使用了 2 个 LRN 层,目的是对局部神经元进行归一化处理。虽然这样可以提高 1.2%~1.4% 的模型泛化能力,但是在移植到 FPGA 上时所占用的计算资源使用率较低,显然,这不符合第 1 章提出

的规整化方法中的规整性要求. 有很多研究验证了 LRN 层对模型性能并没有很大作用, 例如 VGG-Net, 故本文考虑将其删除, 并期望删除后的模型架构满足规整化方法中的运算速度要求.

本文对仅删除 LRN 层的模型在 voc2007 数据集上重训练 5 次, 统计每次精度和速度的结果, 其平均检测结果如表 1 所示. 对删除 LRN 层模型精度 (mAP) 结果进行 t 检验, 结果 $P>0.05$, 按照 $\alpha=0.05$ 统计显著水平, 证明使用 LRN 层和删除 LRN 层精度结果差别无统计学意义.

表 1 LRN vs No_LRN 结果对比

Method	mAP(%)	FPS onPC
LRN	63.4	38.1
No_LRN	62.9	39.6

从表 1 中可以看出, 使用 LRN 层模型相比于删除 LRN 层模型没有带来性能上的提升, 甚至导致内存的消耗和计算时间的增加. 因此选择删除 LRN 层.

3.2.3 替换 Leaky-ReLU 激活函数

如图 1(a) 所示, YOLO-V1 算法神经网络结构中只使用了 1 个 Leaky-ReLU 激活函数, 其余卷积层均使用 ReLU 层, 这不符合规整化方法中规整性要求和低运算复杂度要求. 故在前两个网络结构改进的基础上, 进一步将最后一个 Conv 层的 Leaky-ReLU 激活函数替换为 ReLU.

Leaky-ReLU 激活函数在 $x<0$ 时 $f(x)=\alpha x$. 其中 α 默认值为 0.01 的系数, 可以保证输出小于 0 的数据也会进行很小幅度的更新. 但由于其导数总是不为零, 使得计算量激增, 从而导致整个网络计算速度变慢. 故将 Leaky-ReLU 替换为 ReLU.

4 实验验证与结果分析

4.1 数据集与实验环境

为验证本文改进模型结构的有效性, 本文在 voc2007 数据集上进行实验. voc2007 数据集共包含 20 类目标, 其中训练集图片共 2 501 张, 目标 6 301 个; 测试集图片共 4 952 张, 目标 12 032 个.

实验 PC 平台为搭载了 64 位 Ubuntu16.04 操作系统的计算机, 显卡型号为 NVIDIA GTX 2080Ti. 编程语言为 Python, 并使用 Caffe 深度学习框架编写代码.

实验 FPGA 平台为 Xilinx 的 ZCU102 开发板. 它是基于 Zynq[®] UltraScale + TMXCZU9EG-2FFVB1 156EMPS0 的通用快速评估板, 其功耗为 30 w, 具有很强的实用价值. 它拥有用于定制设计

的 FPGA 逻辑, 这为目标检测算法提供了灵活的开发平台. 并且在平台上使用 CHaiDNN, 它是一个 Xilinx 深度神经网络库.

4.2 评价标准

实验中, 当检测结果与真值边界框的 IOU 超过 0.5 即可认为检测正确. 采用准确率 (precision)、召回率 (recall)、平均准确率 (mean Average Precision, mAP) 作为算法评价指标. mAP 的定义为

$$mAP = \int_0^1 P(R) dR$$

(8)

式中, P 为准确率, R 为召回率. mAP 指标能够更全面地衡量检测算法在多召回率下的精度, mAP 值越大, 模型结构性能越好.

此外, 本文还使用了每秒帧率 (Frames Per Second, FPS) 作为模型结构的另一个性能指标. FPS 值越大, 模型结构性能越好.

4.3 实验结果与分析

4.3.1 在 PC 端的测试

为了验证本文改进的目标检测算法模型结构的有效性, 通过 PC 端将 YOLO-V1 算法与本文改进的算法进行对比, 其对比结果如表 2 所示. 需要注意的是, 本文改进的算法的检测结果是重训练模型后得到. 在训练期间的具体参数设置如下: 动量 (momentum) 为 0.9, 权值衰减系数 (weight_decay) 为 0.000 5, 批量大小 (batch_size) 为 16, 学习率 (base_lr) 为 0.001, 迭代次数 (test_iter) 为 6 000.

表 2 YOLOv1 算法与提出的算法的对比

Method	mAP(%)	FPS onPC
YOLOv1	63.4	38.1
Improved Arithmetic	63.1	42.5

从表 2 可以得出, 本文改进的算法在 PC 端相比传统的 YOLO-V1 算法而言在精度几乎不损失的情况下, 速度大幅度提升, 提升了 11.5%.

4.3.2 在 FPGA 上的测试

传统的 YOLO-V1 算法由于不符合本文提出的规整化理论, 将其加载到 FPGA 平台工作量巨大. 故本文在 FPGA 上只对改进的目标检测模型进行测试, 验证其是否满足高实时性的应用需求. 需要注意的是, 为了更好的利用 FPGA 规整化运算性能、提高检测速度, 本文对改进的模型采用 8 位量化, 在一定程度上算法精度存在少量损失.

从表 3 中可以看出, 本文改进的目标检测算法在精度满足目标检测要求的情况下, 在速度上达到 29 FPS, 满足实时需求.

表 3 提出的算法精度和速度

Method	mAP(%)	FPS on FPGA
Algorithm_proposed	62.3	29

4.3.3 可视化结果显示

图 3(a)表示原图,3(b)表示 YOLO-V1 结果图,3(c)表示改进的目标检测算法在 PC 端的部分可视化检测结果.图中顶部蓝色 Time 表示每张图片的检测速度,红色框表示每个目标的边界,每个红色框上的绿色字体表示预测的目标类别.

从图 3(c)中可以看出,本文改进的检测算法对目标有较高的检测速度,检测框能够接近准确地包围目标,分离目标与背景.检测结果表明模型能够适用于多场景、多目标下的目标检测.

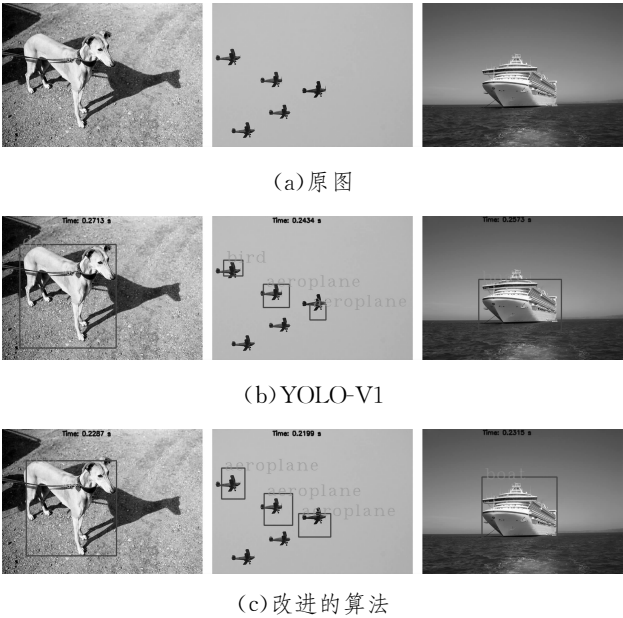


图 3 可视化测试结果

5 结束语

本文改进了一种基于 FPGA 的高速度卷积神经网络目标检测算法,该算法的规整化结构设计,使得检测算法的整体计算速度得到很大提升.将此目标检测算法在 voc2007 数据集上进行了验证.验证结果证明,在 PC 端,本文改进的算法较传统 YOLO-V1 算法而言在精度几乎不损失的情况下,速度提升了 11.5%;在功耗 30w 的 FPGA 上,算法 mAP 达到 62.3%,速度达到 29FPS.算法除了满足实际检测的精度要求外,还极大满足了实时性应用需求.但算法整体也存在一定漏检、误检情况.如何优化特征表征、提升模型在复杂场景图像中的检测精度是后续工作的重点研究方向.

参考文献:

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158. DOI: 10.1109/TPAMI.2015.2437384.

[2] HAO Z K, LIU Y, QIN H W, et al. Scale-aware face detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1913-1922. DOI: 10.1109/CVPR.2017.207.

[3] JIANG Z Q, HUYNH D Q. Multiple pedestrian tracking from monocular videos in an interacting multiple model framework[J]. IEEE Transactions on Image Processing, 2018, 27(3): 1361-1375. DOI: 10.1109/TIP.2017.2779856.

[4] 陈辰,严伟,夏珺,等.基于FPGA的深度学习目标检测系统的设计与实现[J].电子技术应用,2019,45(8): 40-43. DOI: 10.16157/j.issn.0258-7998.190318.

CHEN C, YAN W, XIA J, et al. Design and implementation of FPGA-based deep learning object detection system[J]. Application of Electronic Technique, 2019, 45(8): 40-43. DOI: 10.16157/j.issn.0258-7998.190318.

[5] 方睿,刘加贺,薛志辉,等.卷积神经网络的FPGA并行加速方案设计[J].计算机工程与应用,2015,51(8): 32-36. DOI: 10.3778/j.issn.1002-8331.1405-0335.

FANG R, LIU J H, XUE Z H, et al. FPGA-based design for convolution neural network[J]. Computer Engineering and Applications, 2015, 51(8): 32-36. DOI: 10.3778/j.issn.1002-8331.1405-0335.

[6] QIU J T, WANG J, YAO S, et al. Going deeper with embedded FPGA platform for convolutional neural network[C]//Proceedings of 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays. Monterey: ACM, 2016: 26-35. DOI: 10.1145/2847263.2847265.

[7] 王巍,周凯利,王伊昌,等.卷积神经网络(CNN)算法的FPGA并行结构设计[J].微电子学与计算机,2019,36(4): 57-62. DOI: 10.19304/j.cnki.issn1000-7180.2019.04.012.

WANG W, ZHOU K L, WANG Y C, et al. FPGA parallel structure design of Convolutional Neural Network (CNN) algorithm[J]. Microelectronics & Computer, 2019, 36(4): 57-62. DOI: 10.19304/j.cnki.

issn1000-7180, 2019, 04, 012.

- [8] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv: 1409.1556, 2015.
 - [9] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6517-6525. DOI: 10.1109/CVPR.2017.690.
-
- (上接第 86 页)
- [5] 陈凌翔, 柯永振, 孙玉婷. 一种用于图片来源检测的相机参考模式噪声提取算法[J]. 计算机工程, 2017, 43(12): 227-230. DOI: 10.3969/j.issn.1000-3428.2017.12.041.
CHEN L X, KE Y Z, SUN Y T. A camera reference pattern noise extraction algorithm for image source detection[J]. Computer Engineering, 2017, 43(12): 227-230. DOI: 10.3969/j.issn.1000-3428.2017.12.041.
 - [6] FERREIRA A, CHEN H, LI B, et al. An inception-based data-driven ensemble approach to camera model identification[C]//Proceedings of 2018 IEEE International Workshop on Information Forensics and Security. Hong Kong, China: IEEE, 2018: 1-7. DOI: 10.1109/WIFS.2018.8630774.
 - [7] MAYER O, STAMM M C. Learned forensic source similarity for unknown camera models[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Calgary, Canada: IEEE, 2018: 2012-2016. DOI: 10.1109/ICASSP.2018.8462585.
 - [8] ZHAO M N, WANG B, WEI F, et al. Source camera identification based on coupling coding and adaptive filter[J]. IEEEAccess, 2020, 8: 54431-54440. DOI: 10.1109/ACCESS.2019.2959627.

作者简介:

程钰清 女, (1995-), 硕士研究生. 研究方向为计算机视觉.
贺占庄(通讯作者) 男, (1962-), 研究员. 研究方向为航天嵌入式系统结构设计. E-mail: hzz771@163.com
马 钟 男, (1971-), 博士, 高级工程师. 研究方向为计算机视觉.
毕瑞星 男, (1989-), 硕士, 工程师. 研究方向为计算机视觉.
毛远宏 男, (1980-), 博士研究生. 研究方向为计算机视觉.

- [9] AHMED F, KHELIFI F, LAWGALY A, et al. Comparative analysis of a deep convolutional neural network for source camera identification[C]//Proceedings of the 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability. London: IEEE, 2019: 1-6. DOI: 10.1109/ICGS3.2019.8688260.
- [10] AN Y, WANG B, ZHAO M J, et al. Patch-based sensor pattern noise for camera source identification [C]//Proceedings of 2015 IEEE China Summit and International Conference on Signal and Information Processing. Chengdu: IEEE, 2015: 866-870. DOI: 10.1109/ChinaSIP.2015.7230528.
- [11] GLOE T, BÖHME R. The 'Dresden Image Database' for benchmarking digital image forensics [C]//Proceedings of the 2010 ACM Symposium on Applied Computing. Sierre, Switzerland: ACM, 2010: 1584-1590. DOI: 10.1145/1774088.1774427.

作者简介:

方梦华 女, (1995-), 硕士研究生. 研究方向为图像取证、云计算与信息安全. E-mail: 13151999695@163.com
林果园 男, (1975-), 博士, 副教授. 研究方向为云计算, 信息安全等.