

引用格式:张鹏举,严佩敏,朱秋煜. 基于改进型 FaceBoxes 的人头检测算法[J]. 微电子学与计算机, 2021, 38(1): 33-37.
[ZHANG P J, YAN P M, ZHU Q Y. Head detection algorithm based on improved FaceBoxes [J]. Microelectronics & Computer, 2021, 38(1): 33-37.]

基于改进型 FaceBoxes 的人头检测算法

张鹏举,严佩敏,朱秋煜

(上海大学 通信与信息学院, 上海 200444)

摘 要: 对于监控视频下的垂直视角,常常会出现人体遮挡,背景模糊等导致检测人体精度低的问题,本文提出以人头为目标代替人体进行检测的监控视频人数统计的方法.该方法对一种轻量型的网络 FaceBoxes 进行改进,并加入多尺度特征融合的方式,提升密集人头检测的精度.由于监控视频中,人头目标比较小,一般不会出现特大目标,因此使用 k-means 方法对候选框的尺度进行聚类,再进行训练,并使用交叠区域检测方法提高预测框的精度.实验结果表明,改进型的 FaceBoxes 方法相比之前的网络精度提高了约 4%,其速度稍有降低.

关键词: FaceBoxes; 人头检测; 神经网络; 特征融合

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2021)01-0033-05

Head detection algorithm based on improved FaceBoxes

ZHANG Peng-ju, YAN Pei-min, ZHU Qiu-yu

(School of communication and information, Shanghai University, Shanghai 200444, China)

Abstract: For the vertical perspective of surveillance video, human body occlusion and background blur often occur, which lead to low accuracy of human body detection. This paper proposes a surveillance video number statistics method that takes human head as the target to detect human body. This method improves a lightweight network of FaceBoxes and adds multi-scale feature fusion to improve the accuracy of dense head detection. In the surveillance video, the head target is relatively small, and generally there is no large target. Therefore, k-means method is used to cluster the scale of candidate frame, and then to train, and overlapping region detection method is used to improve the accuracy of prediction frame. The experimental results show that the improved FaceBoxes method is about 4% more accurate and slightly slower than the previous one.

Key words: FaceBoxes; head detection; neural network; feature fusion

1 引言

随着科技的发展,视频监控越来越普遍,大部分视频监控都是为了安全防护而设置的.但视频中大量的信息,却不得而知.对于商铺或者景区,可以通过对视频中的人流进行检测,从而得到各个时间段人流信息,以便做出相应的商业决策或者对于突发情况导致的拥挤现象进行预警作用.现有的人流计数方法主要分为两大类,一种是基于人群热力图的方式来对密集人群的密度与聚集程度来进行预测,另一种基于目标检测的方法对人群个体进行精确定

位.前者可以对高度密集的人群进行估算,但不可以给出其精确定位,以及出入状态.对于一般的特定场合的人流统计,一般会有序进出,因此基于检测的方式更为合适.

目前,人头检测的方案主要有两种.一种是基于传统图像处理的方法,通过人工设计特征,比如梯度直方图、图像深度信息或者 Harr 描述子等,并利用机器学习常用的分类器如支持向量机(SVM)、自适应提升树(AdaBoost)进行分类.但其设计特征复杂,且泛化能力不强,不能适应多个任务.另一种是

基于深度学习技术将图像通过卷积神经网络得到其目标的坐标点与类别^[1]. 其中 RCNN^[2] 首次将自动生成的图像特征应用到目标检测领域中, FPN^[3] 提出了特征融合的方案, 融合高层的语义特征与底层的细节特征, 这些都属于双阶段算法, 其训练复杂, 速度较慢, 但精度较高. 而 SSD^[4] 与 YOLO^[5] 系列使用回归的方法训练网络, 直接在检测过程中生成预选框, 属于单阶段算法, 其训练简单, 速度较快. 近些年, 大部分研究人员为了实际应用模型, 都选择单阶段的方式进行训练, 也出现了一种 Anchor-Free 的方法, 来进一步加快速度和提高精度, 寻找一个速度与精度的平衡状态.

本文使用一种改进型的轻量型网络 FaceBoxes^[6] 用于人头的检测, 更加适用于工业领域嵌入式设备中运行. 本文的研究工作主要有以下几点:

(1) 对 FaceBoxes 网络三个目标检测输出特征层进行融合, 通过对低尺度特征进行上采样后与高尺度特征进行相加或者降维后拼接, 本文使用降维后拼接, 可以更好地混合两个特征层, 提高检测的精度.

(2) 由于人头区域的比例接近 1:1, 而且人头在视频监控的形变也不大, 根据感受野的特点, 采用 K-means++ 对数据集中的真实框进行尺度聚类, 根据 IOU 确定聚类个数, 并设置预选框, 提高检测精度和速度. 相比原论文的方法, 检测精度和速度均有提高.

(3) 对于单阶段算法, 其输入图像长宽比例一般为 1:1, 但一般监控区域为长方形, 因此, 在测试时, 可以通过将测试图像分成两个正方形进行检测, 中间交叠区域使用 NMS 去重, 从而大大提高检测精度.

2 FaceBoxes 检测算法

FaceBoxes 是一种高效轻量化的检测网络, 为了解决密集人脸识别而提出来, 是一种可以在 CPU 上高效运行的目标检测网络, 该模型的结构主要分为两中, 快速消化卷积层 (RDCL) 和多尺度卷积层 (MSCL).

RDCL 通过使用大尺度的卷积核来快速的降低输入图像的尺寸, 减少输出通道数. 为了保证速度与精度的平衡, 对 Conv1、Conv2 以及所有的 Pool 层分别选择 7×7 、 5×5 和 3×3 的卷积核大小, 并且对应的步长分别为 4, 2, 2. 经过 RDCL 后, 总步长为 32, 从而可以快速的降低输入图像的尺寸, 在保持精

度的同时提高网络的推理能力. 为了进一步提高网络运行速度, 使用 CReLU 来减少底层网络的卷积核数量.

MSCL 学习了 Inception 模块, 在卷积层之间加入多个分支, 增加网络宽度, 在同一特征尺度下丰富感受野的范围, 提高小目标检测的精度, 并且速度不会降低太多. 并选择 Inception3、Conv3_2、Conv4_2 来作为目标检测层, 使用多尺度检测的方法可以高效的扩大检测范围, 对形变较大的物体有更高的匹配度.

除此以外, 还使用了预选框稠密化的策略, 在底层检测较小目标时, 生成密度更高的预选框, 这是对密集目标十分有效的一种方案. 其定义如下:

$$A_{\text{density}} = A_{\text{scale}} / A_{\text{interval}}$$

式中, 为了更好地检测小目标, 在 Inception3 设置了三个 anchor, 尺度分别为 32×32 、 64×64 、 128×128 . 而该层的间隔为 32, 通过上面公式得到密度分别为 1, 2, 4. 也就是尺度越小的 anchor 越稀疏, 因此需要进行稠密化的举措.

3 改进型 FaceBoxes 网络

虽然 FaceBoxes 网络非常的轻量化, 但对于尺度较小, 且形变范围小的人头, 精度却不太理想. 在监控视频中, 特别是密集人群中, 人头的像素范围一般在 10 像素到 60 像素之间, 而原 FaceBoxes 算法设计的 anchor 尺寸并不合适. 而且由于人头之间的差异性比较小, 在单一的特征层上进行预测, 精度较差. 为了解决以上问题, 提出了改进型 FaceBoxes 网络, 并在图像预测时使用交叠区域检测的方法来提提高识别精度.

3.1 网络结构的改进

为了加快推理速度, 并且通过先验知识得出人头尺度不超过 250 个像素点, 改进型 FaceBoxes 网络舍弃了在高层 Conv4_2 进行检测. 同时, 带来的缺点就是目标高级语义特征的丢失, 为了解决这个问题, 使用一种特征金字塔的多尺度模块来融合高层语义特征和底层细节特征.

首先, 对高层 Conv4_2 进行上采样为 $16 \times 16 \times 256$, 并与 Conv3_2 的 $16 \times 16 \times 256$ 进行拼接得到 $16 \times 16 \times 512$, 为了保持输出通道数与原来不变, 使用 1×1 的卷积核进行降维处理. 改进型 FaceBoxes 网络使用 3 个 3×3 的深度可分离卷积代替 7×7 卷积核, 2 个 3×3 的深度可分离卷积代替 5×5 的卷积核, 并且不在 Conv4_2 层进行检测. 具体的网络

结构如图 1 所示.

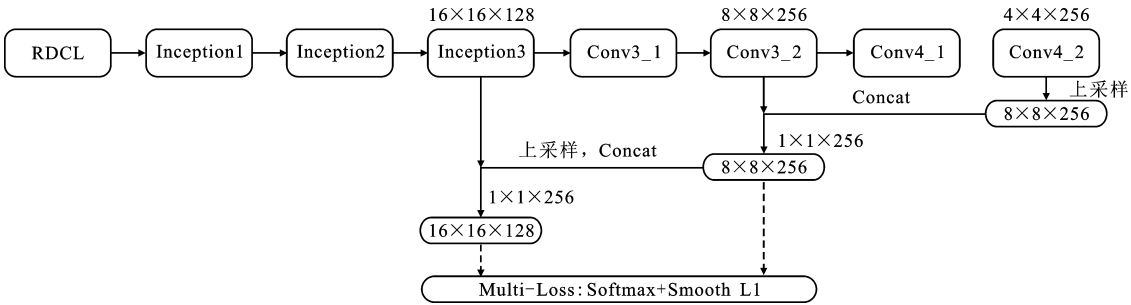


图 1 改进型 FaceBoxes 网络结构图

3.2 使用 K-means 对 Anchor 聚类

在人头数据集中,大部分监控视频中人头不会占据很大的像素,从而原 FaceBoxes 上的 anchor 尺寸并不适合,最小的 anchor 尺寸为 32,会导致很多较小的人头检测不到. 因此需要对人头数据集的真实框重新进行聚类处理.

对于人头检测来说,其目标是一个单一目标,提取到所有真实框的 W 和 H ,并在相对应的图片进行归一化处理,形成聚类数据集 $D = [W_i, H_i]$. 而聚类的目标为每个 anchor 与真实框的 IOU 尽量大,对于神经网络来说本质是梯度下降算法,优化目标是越来越小的,因此可以得到聚类的目标函数 $\sum_{i,j} 1 - \text{IOU}(\text{Anchor}_i, GT_i)$. 其中 i, j 分别表示 anchor 和真实框的标识^[7]. 具体步骤如下:

(1)首先对数据集中的标签进行特征提取,由于卷积神经网络中的卷积运算有滑窗的特质,因此只需要对标签中的 W 和 H 当做特征即可. 但数据集中各个图片尺寸不一致,而且每个图片中目标边界尺度也不同,需要经过归一化处理后才可以进行聚类处理. 可以得到归一化后的数据为 $[W, H]$.

(2)将归一化好的数据使用 Kmeans++ 算法聚类,聚类中心设为 k 个,首先随机生成 k 个聚类中心,然后将样本的中心与类中心重合,计算每个样本距离类中心的距离,距离即为目标函数 $d = 1 - \text{IOU}(\text{Anchor}, GT)$,根据得到的距离将样本分配到各个类别簇中去,每次迭代重新计算中心点并更新类中心.

(3)最终得到 k 个聚类中心,接着乘以网络输入图像尺寸 1 024,即可得到相应的 anchor 尺寸.

(4)按照上面的方法,对人头数据集进行聚类处理后,聚类结果如表 1 所示.

表 1 人头数据集 Kmeans++ 聚类结果

类别数	4 类	5 类
Anchor 尺度	(43,45) (144,174) (88, 104) (63,69)	(60,64) (80,94) (114,132) (160,194) (42,44)
IOU	0.73	0.85

其聚类密度图以及各个样本点的分布情况如图 2.

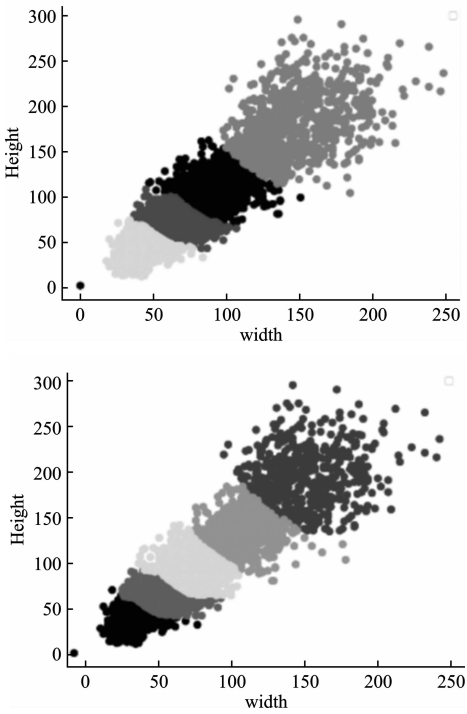


图 2 聚类结果散点分布图(上:4 类,下:5 类)

综上对聚类结果进行分析,考虑使用 5 类更为合适,发现最大尺寸不超过 250,因此可以减少一个尺度层(512x512)的检测任务,只需要在 Inception3 和 Conv_3 这两个特征层上进行检测. 可以加快推理速度. 根据聚类结果调整 anchor 尺寸,为 40x40、60x60、80x80、120x120、160x160,并对

40、80 分别进行 4、2 倍密集化处理,而 120 与 60 不予处理.这样既可以保证速度又可以提高对小人头检测的性能.

由于 Inception3 和 Conv_3 所对应的最小感受野分别为 143×143 、 271×271 ,因此将尺度为 160×160 的 anchor 用于 Conv_3 卷积层中检测,而其余尺度的 anchor 均在 Inception3 卷积层中检测.

3.3 交叠区域检测

由于人头数据集的像素为 $1\,920\times1\,080$,直接无操作送入神经网络中会导致其底层卷积计算量过大,进而影响模型的推理能力.并且由于 FaceBoxes 网络的输入尺寸一般为 $1\,024\times1\,024$,直接对输入图像进行 resize 操作,则会导致人头区域变形影响检测效果.为了更好的检测高像素的人头图片,可以使用交叠区域检测的方式.其具体的操作步骤如下:

(1)将 $1\,920\times1\,080$ 的图片按照左边界和右边界分割成两个 $1\,080\times1\,080$ 的图片,从而造成中间 240 个像素点的交叠区域,而人头最大尺度不超过 250,因此不会出现横跨交叠区域,导致检测失效的情况.

(2)对两个 $1\,080\times1\,080$ 的图片分别送入网络中,得到相应的检测框,并整合到一起.

(3)使用非极大抑制算法对整合后的检测框去重,设置阈值为 0.3.

实验证明,虽然二次检测会降低推理速度,但是精度大大提高,并且与直接输入原图尺寸 $1\,920\times1\,080$ 的方式相比,其速度并没有降低很多.

4 实验与结果分析

4.1 数据集准备和模型训练

由于人头数据集很少且非常老旧和像素质量偏低,本文使用的人头数据集为自建数据集,通过对 50 多段监控行人的视频进行拆分成图片作为人头数据的训练集,共 28 830 张人头图片.包含户外客厅、室内过道等复杂场景以及长发、短发,是否戴帽子等特殊人头图像.并从中挑选 3 000 张图片作为测试集,用于模型性能和泛化能力的衡量.

本文训练环境在 CPU IntelCore i7-7550U、主频为 2.8 GHz、八核、内存为 16 G、Ubuntu16.04 的 Linux 操作系统、显卡为 1 080TI 的实验环境下进行的.由于图片尺寸为 $1\,920\times1\,080$,直接缩放可能会造成人头形变,因此采用随机裁剪的方法,对人头所在区域裁剪成 $1\,024\times1\,024$ 的图像并送

入网络中去,并使用随机水平翻转,随机亮度调整,随机色度调整等数据增强手段^[8-9].在网络训练中,以 32 张图片为一个批次训练 10 个 epoch,共迭代 9 000 次,学习率为 0.001,权重衰减为 $1e-4$,学习率下降策略是分别在 5、7 个 epoch 时学习率降低为原来的 0.1 倍.

4.2 实验结果和理论分析

为了证明本文提出改进方法的有效性,采用平均准确率 mAP、检测时间两个指标在同一条件下对改进前和改进后两个模型进行测试评价.其中 mAP 一般用于检测任务的衡量指标,这里我们采用 pascal voc2012 的评价指标.计算方法是根据精确率 precision 和召回率 recall 所形成的曲线面积计算得来的,定义如下:

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$
$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

式中,TP 为检测框与真实框 $\text{IOU}>0.5$ 的检测框数量,FP 为检测框与真实框 $\text{IOU}\leq0.5$ 的检测框的数量,FN 为没有检测到的真实框的数量.

由于人头检测为单类检测, P_R 为 precision 和 recall 形成的曲线,故其 AP 也就是 mAP,为 P-R 曲线的线下面积,定义如下:

$$\text{AP}=\int_0^1 P_R \text{d}r$$

在 CPU 环境中,测试集上的性能测试结果如表 2 所示.

表 2 改进前后模型性能对照表,*表示使用交叠区域检测方法(CPU 4110@2.1 GHz)

指标	AP(平均准确率)/%	帧数/s ⁻¹
YOLOV3	89.62	/
FaceBoxes	92.57	7
改进后 FaceBoxes	95.18	8
改进后 FaceBoxes *	96.28	4

原 FaceBoxes 和改进后 FaceBoxes 的 P-R 曲线如图 3、4 所示.

选择部分测试样本进行比对,发现改进后 FaceBoxes 的漏检率大大降低,并且在密集人头检测任务中,检测效果远远好于原来 FaceBoxes 网络,对比样本图如图 5.

5 结束语

本文通过对 FaceBoxes 算法进行改进,使其更

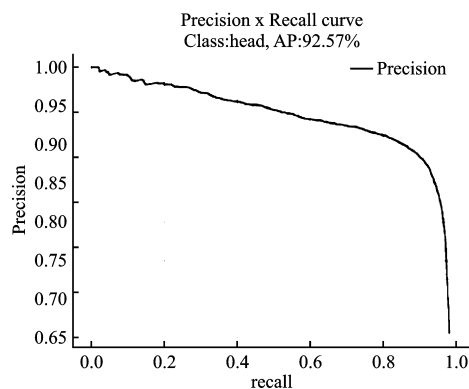


图 3 改进前 P-R 曲线

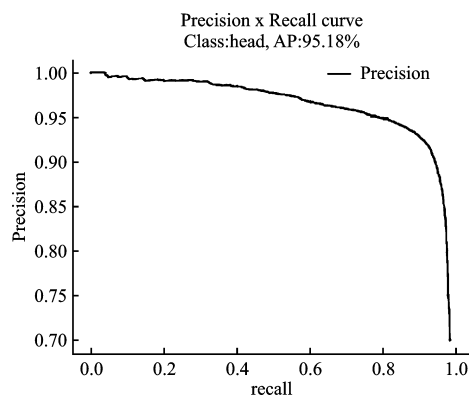


图 4 改进后 P-R 曲线



(a)原 FaceBoxes 检测结果

(b)改进后并交叠检测的
FaceBoxes 检测结果

图 5 对比样本图

适合俯拍监控视频中的人头检测任务. 通过 K-means++ 聚类可以确定更为合适的预选框, 并根据聚类后的预选框进行相应的密集化策略; 根据先验知识只在两个尺度层进行检测, 同时为了加入高层语义特征, 引入特征融合模块. 最后, 在对高像素图片进行检测时引入交叠区域检测方法, 减少模型的漏检率, 提高检测精度, 同时速度不会降低太多.

参考文献:

- [1] 吉训生, 吴凡. 基于混合卷积神经网络的人头检测方法[J]. 高技术通讯, 2018, 28(4): 313-319. DOI: 10.3772/j.issn.1002-0470.2018.04.004.
- JI X S, WU F. Head detection using hybrid convolution neural networks[J]. Chinese High Technology Letters, 2018, 28(4): 313-319. DOI: 10.3772/j.issn.1002-0470.2018.04.004.
- [2] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: ACM, 2015: 91-99.
- [3] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2117-2125. DOI: 10.1109/CVPR.2017.106.
- [4] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//14th European Conference on Computer Vision. Amsterdam: Springer, 2016. DOI: 10.1007/978-3-319-46448-0_2.
- [5] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7263-7271. DOI: 10.1109/CVPR.2017.690.
- [6] ZHANG S F, ZHU X Y, LEI Z, et al. FaceBoxes: a CPU real-time face detector with high accuracy[C]//2017 IEEE International Joint Conference on Biometrics (IJCB). Denver: IEEE, 2017: 1-9. DOI: 10.1109/BTAS.2017.8272675.
- [7] 高玮军, 师阳, 杨杰, 等. 一种改进的轻量人头检测方法[J/OL]. 计算机工程与应用, 2020: 1-9. (2020-02-21). <https://kns.cnki.net/KCMS/detail/11.2127.tp.20200221.0926.008.html>.
- [8] 刘阳, 李银萍. 基于视频分析的复杂场景人数统计方法研究[J]. 信息通信, 2018(5): 49-53. DOI: 10.3969/j.issn.1673-1131.2018.05.021.
- LIU Y, LI Y P. Research on the method of people counting based on video analysis in complex scenes[J]. Information & Communications, 2018(5): 49-53. DOI: 10.3969/j.issn.1673-1131.2018.05.021.

(下转第 44 页)

- IEEE Transactions on Nuclear Science, 2013, 60(3): 1816-1823. DOI: 10.1109/TNS.2013.2247630.
- [2] 张森, 石军, 王九龙. 卫星在轨失效统计分析[J]. 航天器工程, 2010, 19(4): 41-46. DOI: 10.3969/j.issn.1673-8748.2010.04.007.
- ZHANG S, SHI J, WANG J L. Satellite on-board failure statistics and analysis[J]. Spacecraft Engineering, 2010, 19(4): 41-46. DOI: 10.3969/j.issn.1673-8748.2010.04.007.
- [3] 何益百. 辐射效应地面试验技术研究[D]. 长沙: 国防科学技术大学, 2010.
- HE Y B. Research on radiation effects ground-based testing technology[D]. Changsha: National University of Defense Technology, 2010.
- [4] 马英起, 封国强, 上官士鹏, 等. 脉冲激光试验评估模拟电路单粒子效应[J]. 信息与电子工程, 2012, 10(5): 633-638.
- MA Y Q, FENG G Q, SHANGGUAN S P, et al. Pulsed laser test of single event effects in analog circuits[J]. Information and Electronic Engineering, 2012, 10(5): 633-638.
- [5] 余永涛. 脉冲激光模拟 SRAM 单粒子效应的试验研究[D]. 北京: 中国科学院研究生院, 空间科学与应用研究中心, 2015.
- YU Y T. Research of SRAM single event effects based on pulsed laser testing[D]. Beijing: Graduate School of Chinese Academy of Sciences, Center for Space Science and Application Research, 2015.
- [6] ZUZARTE M. A tool for run time soft error fault injection into FPGA circuits[D]. Hamilton: McMaster University, 2014.
- [7] KRETZSCHMA U, ASTARLOA A, JIMÉNEZ J, et al. Compact and fast fault injection system for robustness measurements on SRAM-Based FPGAs[J]. IEEE Transactions on Industrial Electronics, 2014, 61(5): 2493-2503. DOI: 10.1109/TIE.2013.2273476.
- [8] QUINN H, WIRTHLIN M. Validation techniques for fault emulation of SRAM-based FPGAs[J]. IEEE Transactions on Nuclear Science, 2015, 62(4): 1487-1500. DOI: 10.1109/TNS.2015.2456101.
- [9] ZHANG R S, XIAO L Y, LI J, et al. A fast fault injection platform of multiple SEUs for SRAM-based FPGAs[J]. Microelectronics Reliability, 2018, 82: 147-152. DOI: 10.1016/j.microrel.2018.01.014.
- [10] 李鹏. 纳米级 SRAM 单粒子翻转效应及其诱导的软错误研究[D]. 长沙: 国防科学技术大学, 2016.
- LI P. Research on single event upset and its induced soft error of Nano SRAMs[D]. Changsha: National University of Defense Technology, 2016.

作者简介:

王梦茹 女, (1991-), 博士研究生, 工程师. 研究方向为 SRAM 型 FPGA 单粒子故障注入与安全性设计.

E-mail: wmmomo@163.com.

周珊 女, (1983-), 博士研究生, 高级工程师. 研究方向为航天 FPGA/SOPC 验证方法及智能软件测试.

张弛 男, (1990-), 硕士, 工程师. 研究方向为航天软件高可靠软件验证方法及故障诊断.

王金波 男, (1978-), 博士, 研究员. 研究方向为航天软件高可靠软件测试验证技术、高可靠测试系统的信息安全.

(上接第 37 页)

- [9] 吉训生, 王昊. 基于优化可形变区域全卷积神经网络的人头检测方法[J]. 激光与光电子学进展, 2019, 56(14): 121-131. DOI: 10.3788/LOP56.141009.
- JI X S, WANG H. Head detection method based on optimized deformable regional fully convolutional neural networks[J]. Laser & Optoelectronics Progress, 2019, 56(14): 121-131. DOI: 10.3788/LOP56.141009.

作者简介:

张鹏举 男, (1997-), 硕士研究生. 研究方向为小目标检测、图像分类、人脸检测与识别.

严佩敏 女, (1963-), 博士, 副教授. 研究方向为智能信息系统、图像处理、电路与系统.

朱秋煜(通信作者) 男, (1964-), 博士, 教授. 研究方向为计算机视觉、图像处理、模式识别、智慧城市.

E-mail: zhuqiuyu@staff.shu.edu.cn.