

引用格式:李斌,许华杰,吴朝晖. 基于 FPGA 的多人脸实时检测系统设计与实现[J]. 微电子学与计算机, 2021, 38(4): 57-62.
[LI B, XU H J, WU Z H. Design and implementation of FPGA-based real time multiple face detection system[J]. Microelec-
tronics & Computer, 2021, 38(4): 57-62.]

基于 FPGA 的多人脸实时检测系统设计与实现

李 斌^{1,2}, 许华杰^{1,2}, 吴朝晖^{1,2}

(1 华南理工大学 微电子学院, 广东 广州, 510641,

2 人工智能与数字经济广东省实验室(广州), 广东 广州 510641)

摘 要: 设计了一种基于移动端深度学习目标检测算法 MobileNetV2-SSDlite 的小规模、高精度的人脸检测算法。在算法基础上, 围绕硬件架构、片上缓存、性能及功耗等方面进行移动端专用型神经网络加速器的研究与设计; 在 ALINX AX7350 SOC 平台上实现了实时多人脸检测加速器系统。结果表明, 本系统在 100 MHz 时钟下, 平均计算性能为 56.01 GOPS, 功耗为 7.3 W。与现有的 MobileNetV2-SSDlite 加速器相比, 运算速度提高了 28.7%, 资源平均节省了 44.46%, 功耗降低了 26.3%。在 224×224 分辨率下, 达到了 83.4 FPS, 满足实时性需求。

关键词: 人脸检测; MobileNetV2 算法; 硬件加速; FPGA

中图分类号: TN492

文献标识码: A

文章编号: 1000-7180(2021)04-0057-06

Design and implementation of FPGA-based real time multiple face detection system

LI Bin^{1,2}, XU Hua-jie^{1,2}, WU Zhao-hui^{1,2}

(1 School of Electronic and Information Engineering, South China University of Technology,

Guangzhou 510640, Guangdong China; 2 Artificial Intelligence and Digital Economy

Guangdong Laboratory (Guangzhou), Guangzhou 510641, Guangdong China)

Abstract: A small and high-precision face detection algorithm based on MobileNetV2-SSDlite is designed. To improve the computing capability of mobile terminals, the study of the mobile algorithm specific neural network accelerator is carried out in terms of hardware architecture, on-chip memory, performance and power consumption. A real-time multi-face detection accelerator system is simulated and implemented on ALINX AX7350 SOC platform. The results show that under a 100MHz clock, the face detection system has an average computing performance of 56.01 GOPS and a power consumption of 7.3W. Compared with the existing MobileNetV2-SSDlite accelerator, the operation speed is improved by 28.7%, the resources are reduced by 44.46%, and the power consumption is reduced by 26.3%. The speed of the system in inferencing a 224×224 resolution picture reached 83.4 FPS, meet the demand of real-time performance.

Key words: face detection; MobileNetV2 Network; hardware accelerator; FPGA

收稿日期: 2020-05-28; 修回日期: 2020-07-03

基金项目: 国家自然科学基金资助(61571196); 广东省重点领域研发计划项目(2019B010145001); 中央高校基本科研业务费专项资金资助(D2190010)

1 引言

人脸检测作为计算机视觉中的一项重要研究课题,是人脸识别、人脸跟踪、人脸确认、人脸修正等研究的基础工作^[1].传统的人脸检测算法误检率高、召回率低.在复杂的环境下,基于卷积神经网络的人脸检测能够提高检测率,面对多人脸多角度有着出色的效果.为了充分利用边缘算力,减少云端计算的负荷及传输延时,需要设计高精度的轻量化人脸检测模型.

目前基于卷积神经网络(CNN)的人脸检测的应用方式是主机训练,终端推理(Inference).CNN的特点是计算量大、操作重复性高、对算力及资源要求高,难以满足自动驾驶、视频监控等端侧算力有限的实时场景的应用需求.随着网络层数不断加深,移动 CPU 算力不足,GPU 的通用性高但能效比低,研究者开始转向对 FPGA 及 ASIC 平台的 CNN 加速器研究. FPGA 与 ASIC 相比灵活性较高,研发成本及周期较短,可随模型快速更新迭代.

本文依据实时视频监控的应用场景,围绕谷歌移动端框架 MobileNetV2 进行轻量化人脸检测算法设计.优化算法人脸锚定盒子的数量及尺寸,并在 WIDER FACE 数据集上进行训练及评估.在此基础上,提出专用卷积神经加速器结构,在 ALINX AX7350 SOC 平台上对实时多人脸检测系统进行仿真及实现.

2 移动端检测框架 MobileNet

2017 年,谷歌发布了第一代移动端网络框架 MobileNet^[2].与标准的空间卷积方式不同,MobileNet 的卷积核心可以分为两部分,分别是深度(Depthwise)卷积和逐点(Pointwise)卷积. MobileNet 在执行 Depthwise(DW)卷积时把收集每个输入通道的特征,通过 Pointwise(PW)把每个通道的各个点的特征组合起来,此卷积方式能够大大减少计算量,适合用于端侧推理.

2018 年,第二代 MobileNetV2 在第一代 MobileNet 的深度可分离卷积核心基础上增加线性瓶颈结构(Linear bottleneck)^[3].在前一代 MobileNet 中,深度卷积输出通道数与输入通道数相同,无法改变通道数.如图 1 所示,在通道数少的情况下无法获取高维度的特征,因此 MobileNetV2 在进行 DW 卷积前首先进行一次 PW 卷积扩张通道数,接

着在高维度空间进行 DW 卷积,最后再进行一次 PW 卷积降维,并经过激活函数后输出.另外,收到 ResNet 的启发,在不同层间引入了短接(Shortcut)操作. Shortcut 操作能够复用浅层特征,加快模型的收敛速度,使神经网络往更深层发展成为可能.这种先升维后降维,最后经过 Shortcut 成为 MobileNetV2 的基本 Linear bottleneck(LBN)倒残差结构.

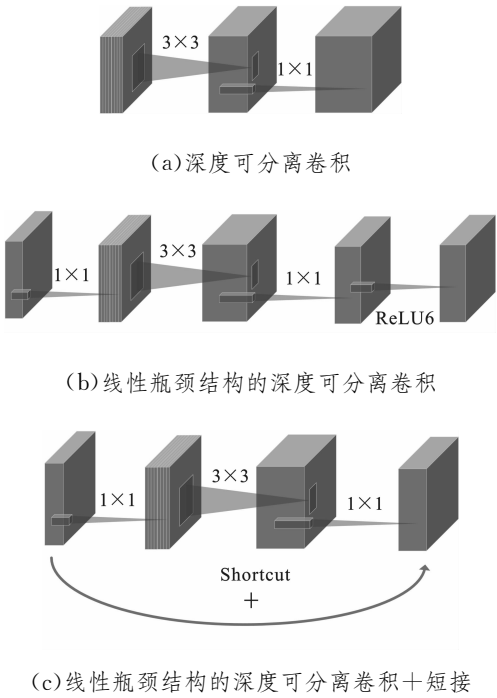


图 1 MobileNetV2 卷积结构示意图

3 移动端人脸检测网络设计

3.1 MobileNetV2-SSD 网络设计

本文根据 MobileNetV2-SSDLite 网络,将锚定盒子先验框从原有的矩形改为方形,同时每级从第一层检测层开始,锚定盒子大小按照还原为原图大小的框大小的 1/4 设置步长.每级检测层的锚定盒子大小设为前一层的 2 倍.改善后的 SSDLite 能更好的匹配人脸的感受野,同时 MobileNetV2-SSDLite 与 VGG-SSD300 相比参数量仅为 其 11.91%^[3],使移动端推理的参数进一步减少.网络结构如表 1 所示.其中 t 为 PW 卷积升维输入通道数的倍数, c 为 LBN 结构的输出通道数, n 为该结构的重复次数, s 为卷积的步长.表 2、表 3 分别为单个 LBN 及 SSDLite 部分 DW 卷积和 PW 卷积结构,其中 k 为输入通道数, k^* 为输出通道数,卷积步长为 s .

表 1 MobileNetV2-SFD 网络结构图

输入	操作	t	c	n	s
$224^2 \times 3$	标准卷积	—	32	1	2
$112^2 \times 32$	LBN	1	16	1	1
$112^2 \times 16$	LBN	6	24	2	2
$56^2 \times 24$	LBN	6	32	3	2
$28^2 \times 32$	LBN	6	64	4	2
$14^2 \times 64$	LBN	6	96	3	1
$14^2 \times 96$	LBN	6	160	3	2
$7^2 \times 160$	LBN	6	320	1	1
$7^2 \times 320$	LBN	—	1280	1	1
$7^2 \times 1\ 280$	标准卷积	—	—	1	—
$7^2 \times 1\ 280$	LBN	0.2	512	1	1
$7^2 \times 512$	LBN	0.25	256	1	1
$7^2 \times 256$	LBN	0.5	256	1	1
$14^2 \times 576$	DW+PW	—	12	1	1
$14^2 \times 576$	DW+PW	—	6	1	1
$7^2 \times 1\ 280$	DW+PW	—	24	1	1
$7^2 \times 1\ 280$	DW+PW	—	12	1	1
$7^2 \times 512$	DW+PW	—	24	1	1
$7^2 \times 512$	DW+PW	—	12	1	1
$7^2 \times 256$	DW+PW	—	24	1	1
$7^2 \times 256$	DW+PW	—	12	1	1
$7^2 \times 256$	DW+PW	—	24	1	1
$7^2 \times 256$	DW+PW	—	12	1	1

表 2 Linear bottleneck (LBN)操作

输入	操作	输出
$h \times w \times k$	1×1 2D 卷积, ReLU6	$h \times w \times tk$
$h \times w \times tk$	3×3 DW 卷积, 步长为 s, ReLU6	$(h/s) \times (w/s) \times tk$
$(h/s) \times (w/s) \times tk$	线性 1×1 卷积, ReLU6	$(h/s) \times (w/s) \times tk^*$

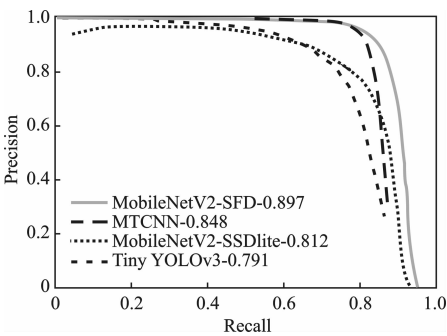
表 3 DW+PW 卷积操作

输入	操作	输出
$h \times w \times tk$	3×3 Depthwise 卷积, 步长为 s, ReLU6	$(h/s) \times (w/s) \times tk$
$(h/s) \times (w/s) \times tk$	线性 1×1 卷积, ReLU6	$(h/s) \times (w/s) \times tk^*$

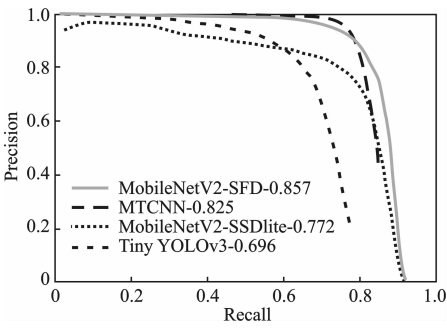
网络采用 WIDER FACE 数据集^[4]进行训练. WIDER FACE 数据集是由香港中文大学收集的权威人脸数据集. 该数据集有着32 203张图片, 总共包含393 703张在不同尺度比例、不同光线条件、以及不同遮挡不同姿势的人脸标签. 为了解决线性瓶颈网络层数过多难以收敛的问题, 本文去除 WIDER FACE 的训练集中小于 10×10 像素的人脸. 为了使人脸的尺度不变, 尽可能覆盖大的人脸, 将训练的输入图统一按 640×640 大小等比例切割.

训练批次大小设为 16, IOU 设为 0.3, 基础学习率设为0.000 5, 动量设为 0.9. 学习率从 0.001 开始, 80 000次迭代后下降至0.000 1, 100 000次迭代后下降至0.000 01, 最大迭代次数设为120 000次.

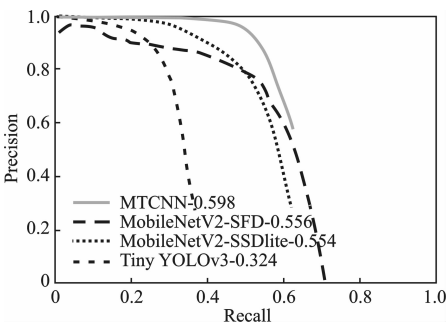
图 2 为 WIDER FACE 评估集算法 PR 曲线. 从图 2 可知, 本设计 MobileNetV2-SFD 算法与 MobileNetV2-SSDlite 相比, 平均精度在简单、中等、困难样本中分别提高 0.085、0.085、0.002. 与 TinyYOLOv3^[5]相比, 平均精度分别提高 0.106、0.161、0.232. 与 MTCNN^[6]算法相比, 分别提高 0.049、0.032, 但在困难样本下, 降低了 0.042. MTCNN 算法推理速度随着人脸数增加而降低, 且检测速度比本算法慢. 因此, 本算法更适用于视频监控等实时多人脸检测场景.



(a) 简单样本



(b) 中等样本



(c) 困难样本

图 2 WIDER FACE 评估集算法 PR 曲线对比

3.2 人脸检测网络量化

为了进一步减少移动端的带宽、降低计算的功耗,本文对设计的算法过程中的浮点操作进行 8-bit 全定点量化^[7]。如表 4 所示, MobileNetV2-SFD224-8bit 为在 224×224 分辨率下 8-bit 全定点量化测试结果。量化后平均精度在简单、中等、困难样本中有所下降,分别为 0.825、0.717、0.338。与原始大小输入图片的量化前 MobileNetV2-SFD 相比,简单样本降低 0.072,中等降低 0.14,困难降低 0.218。其中,困难样本的精度下降较为明显,一方面是因为量化的误差随着网络层数加深不断积累,另一方面是困难样本中人脸的尺寸比较小,把图片分辨率缩小为 224×224 进行实际测试后,小脸检测率降低,经过后处理层量化后人脸框位置也会存在一定的偏移误差。

表 4 本算法量化后 WIDER FACE 评估集对比

输入	简单样本	中等样本	困难样本
MobileNetV2-SFD	0.897	0.857	0.554
MobileNetV2-SFD224-8bit	0.825	0.717	0.338

4 硬件系统设计

4.1 系统整体描述

系统采用 ALINX AX7350 SOC 平台开发。如图 3 所示,处理器(PS)通过 AXI 总线传递卷积需要的参数、指令。AXI4 总线完成摄像头数据从 PS 端 DDR 至 PL 端搬移,可编程逻辑(PL)部分通过 AXI 交互连接器 (AXI Interconnect) 和直接内存控制器 (SDMA) 与 PL 端 DDR 交互,完成人脸检测网络的推理。最后结果通过 AXI 总线传回 PS,PS 接受 PL 完成的一帧图像的中断后在屏幕上显示。

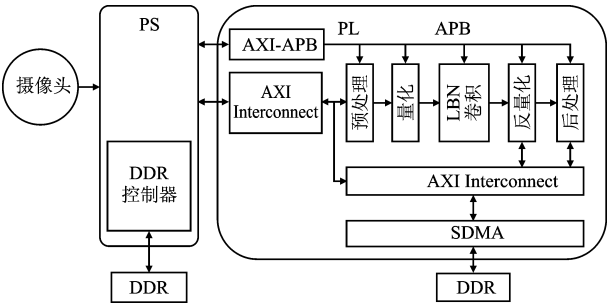


图 3 整体硬件方案框图

4.2 线性瓶颈结构加速单元

针对 MobileNetV2 中的特殊结构,本文以其中的线性瓶颈(LBN)结构作为运算的基本单元,完成 PW 卷积升维、DW 卷积提取通道特征、PW 卷积降

维、短接及激活函数(RELU/RELU6)的操作。DW 卷积无法复用输入特征图,对外部 DDR 带宽需求极大。升维和降维的输入,输出通道数较少,且为 1×1 的卷积参数量较少。因此本文根据 LBN 结构,把三层合并为一层进行运算把升维后的中间结果缓存在带宽较大的内部 BRAM 中,解决片外内存带宽不足影响计算效率的问题。如图 4 所示,LBN 结构首层普通卷积和 PW 卷积合并在一个数据流计算,DW 卷积单独一个数据流计算。PW 和普通卷积中间结果存放至 32-bit PW 缓存,DW 卷积结果经反量化后存放至 8-bit DW 缓存。以 LBN 结构为基本单元,完成三层卷积操作和短接(Shortcut)操作后输出到 DDR。在输入特征图间并行中设计可配置的加法树,适配两种不同的卷积运算。另外在卷积时,将 32 位的浮点数量化为 8 位,进一步降低内存和带宽。除此之外,对输入特征图、权值偏置以及量化参数采用乒乓缓存的方式,相互交替数据缓存和工作缓存,减少加载数据的等待时间。后处理部分经过量化后查表完成坐标还原及 SoftMax,最后采用对内存消耗较少的归并排序方案,全程流水线地在 PL 部分进行以降低延迟。

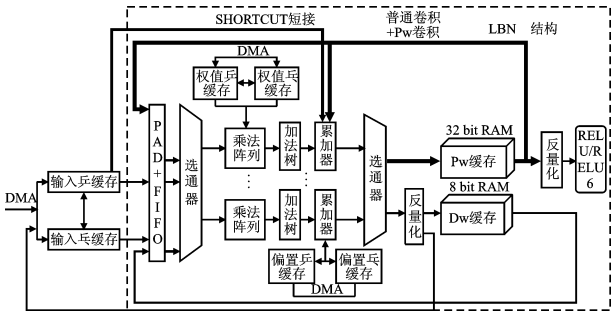


图 4 LBN 结构数据流框图

5 实验结果与分析

本设计在综合时使用 LUT 进行 8-bit 乘法单元综合。分析时钟路径后,发现时钟瓶颈在加法树,因此用速度较快的 DSP 综合加法树提高系统频率。

表 5 推理加速器与 CPU 性能对比

平台	类型	频率(Hz)	推理时间(ms)	功耗(W)	能效比(GOPS/W)
RTX 2070S	GPU	1.6G	5.57	85	1.42
Ryzen 3700X	CPU	3.6G	103.64	65	0.10
Kirin 970	CPU	2.4G	34.29	4.6	4.25
本设计	FPGA	100M	11.98	7.3	7.67

本节进行移动端边缘神经网络加速器的推理性

能对比,首先与通用 GPU、CPU 及移动 MCU 进行对比.如表 5 所示,在 224×224 分辨率下,本网络的整体乘累加数为 671.59 M,推理性能为通用 CPU 的 8.65 倍.在移动端麒麟(Kirin)970 处理器利用腾讯 NCNN 框架进行测试,推理性能加速比为 2.66 倍.推理加速器性能相比 Kirin 970 处理器有着更高的能耗比和计算速度优势本.本设计算力不及 GPU,但是其能耗比为 GPU 的 5.4 倍.

现有的文献中,大多数针对 MobileNetV1-SSD 进行加速器设计,MobileNetV1-SSD 中不具有线性瓶颈结构,运算量较大.文献[8]为 MobileNetV1-SSD 加速器,采用 32 位浮点精度运算.其加法树在进行 DW 卷积时未经过优化,与本文的加速器相比,本文加速器采用 8-bit+32-bit 混合精度,乘法运算综合时尽量使用 LUT 替代 DSP,因此大幅度减少 DSP 使用量. Slice LUT、Slice Register、DSP 和 BRAM 资源与文献[8]相比分别节省了 12.23%、58.8%、92.94%、31.07%的资源.由于网络结构及运算量不同,比较其能效比.本设计能效比为文献[8]的 1.83 倍,具有较高的能效比.

文献[9]为 MobileNetV2-SSDLite 的推理加速器,网络有着相同线性瓶颈结构,该加速器对网络进行了剪枝,减少了其通道数,但网络基本结构不变,精度一致为 8-bit+32-bit 混合精度.本文根据文献[9]中的网络的结构计算其整体的乘累加操作数为 670.31M,所需的乘累加数与本设计的网络几乎一致.另一方面,由于文献[9]中的预处理部分和后处理层均在 ARM 上运行,为了更公平地进行对比,本文采用对比的资源数为去除后处理层的资源数.

表 6 基于 FPGA 推理加速器性能对比

	文献[8]	文献[9]	本设计
平台	Zynq7100	ZC706	AX7350
工艺(nm)	28	28	28
网络算力(MOP)	1.30×10^3	670.31	671.59
计算精度	32-bit	8-bit + 32-bit	8-bit + 32-bit
Slice LUT	142291	148484	124894
Slice Register	187146	191899	77104
DSPs	1926	728	136
BRAM(36K)	354	311	244
频率(MHz)	100	100	100
算力(GOPS)	17.11	43.44	56.01
推理时间(ms)	—	12.43	11.98
帧率(fps)	—	64.8	83.4
功耗(W)	4.1	9.9	7.3
能效比(GOPS/W)	4.17	4.39	7.67

如表 6 所示,在计算量和文献[9]相近的情况下,本文设计的加速器运算速度提高了 28.7%.此外,本文设计的加速器 Slice LUT、Slice Register、DSP 和 BRAM 资源与文献[9]相比分别节省了 15.89%、59.82%、81.32%、21.54%的资源,平均节省了 44.46%的资源,降低了 26.3%的功耗.在输入为 224×224 分辨率情况下,帧率为 83.4 FPS,能效比为其 1.75 倍.本设计在相同精度下,速度上占优且降低了资源及功耗,具有一定的先进性.

6 结束语

本文针对终端人脸检测成本及功耗高的问题,设计了轻量化的人脸检测网络 MobileNetV2-SFD,在 ALINX AX7350 SOC 开发板上实现了轻量化多人脸实时检测系统.结果表明,本系统在 WIDER FACE 评估集简单、中等、困难样本中平均检测精度分别达到 0.825、0.717、0.338.与现有同类型加速器相比,运算速度提高 28.7%,平均节省 44.46%的资源,降低了 26.3%的功耗.在 224×224 分辨率达到了 83.4 FPS.本系统的检测速度和资源消耗与现有的同类型加速器相比,均显示出优势.

参考文献:

[1] 秦琴. 基于肤色和几何特征的人脸检测与识别方法[J]. 科技风, 2018, 31(1): 232—232. DOI: 10.19392/j.cnki.1671-7341.201801208.

QIN Q. Face detection and recognition based on skin color and geometric features[J]. Technology Wind, 2018, 31(1): 232. DOI: 10.19392/j.cnki.1671-7341.201801208.

[2] HOWARD A G, ZHU M L, CHEN B, et al. Mobile-nets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2020-07-07]. <https://arxiv.org/abs/1704.04861>.

[3] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510-4520. DOI: 10.1109/CVPR.2018.00474.

[4] YANG S, LUO P, LOY C, et al. WIDER FACE: a face detection benchmark[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 5525-5533. DOI: 10.1109/CVPR.2016.596.

[5] REDMON J, FARHADI A. YOLOv3: an incremental

- improvement[EB/OL]. (2018-04-08)[2020-07-07].
<https://arxiv.org/abs/1804.02767>.
- [6] ZHANG K P, ZHANG Z P, LI Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10): 1499-1503. DOI: 10.1109/LSP.2016.2603342.
- [7] XU H J, WU Z H, DING J, et al. FPGA based real-time multi-face detection system with convolution neural network[C]//Proceedings of the 2019 8th International Symposium on Next Generation Electronics. Zhengzhou: IEEE, 2019: 1-3. DOI: 10.1109/ISNE.2019.8896551.
- [8] LIU B, ZOU D Y, FENG L, et al. An FPGA-based CNN accelerator integrating depthwise separable convolution[J]. Electronics, 2019, 8(3): 281. DOI: 10.3390/electronics8030281.
- [9] FAN H X, LIU S L, FERIANC M, et al. A real-time object detection accelerator with compressed SSDLite on FPGA[C]//Proceedings of 2018 International Conference on Field-Programmable Technology. Naha: IEEE, 2018: 14-21. DOI: 10.1109/FPT.2018.00014.

作者简介:

李 斌 女,(1967-),博士,教授.研究方向为集成电路设计研究. E-mail: phlibin@scut.edu.cn

许华杰 男,(1994-),硕士研究生.研究方向为数字集成电路设计.

吴朝晖 男,(1971-),博士,副教授.研究方向为集成电路设计研究.