ICS 25.040.30
J28

SZROBOT

深圳市机器人协会团体标准

T/SZROBOT 0005—2023

挂载智能芯片的边缘视觉模组技术要求

Technical requirements for AI-chip enhanced edge vision module

点击此处添加与国际标准一致性程度的标识

征求意见稿

2023/1/30

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

深圳市机器人协会　发布

目 录

前言 III
挂载智能芯片的边缘视觉模组技术规范 1
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 缩略语 2
5 边缘视觉模组技术架构 2
5.1 总体技术架构 3
5.2 硬件层 3
5.3 架构层 3
5.4 工具层 4
5.5 应用层 4
6 边缘智能模组硬件层技术要求 5
6.1 模块意义 5
6.2 功能要求 5
7 边缘智能模组架构层技术要求 6
7.1 图像信号处理器技术要求 6
7.2 张量处理器技术要求 7
7.3 浮点执行单元技术要求 8
7.4 现场可编程阵列技术要求 8
8 边缘智能模组工具层技术要求 9
8.1 模型解析技术要求 9
8.2 模型量化技术要求 9
8.3 模型精度评估技术要求 10
8.4 模型部署技术要求 10
9 边缘智能模组算法层技术要求 11
9.1 目标分类网络技术要求 11
9.2 目标识别网络技术要求 11
9.3 语义分割网络技术要求 11
9.4 超分辨率网络技术要求 12
9.5 图像去噪网络技术要求 12

前言

本标准按照GB/T 1.1-2020给出的规则起草。

本标准由深圳市机器人协会提出并归口。

本标准起草单位：中国科学院深圳先进技术研究院，中国科学院长春精密仪器与物理研究所，西安瑞峰光电有限公司，深圳市辰卓科技有限公司

本标准主要起草人：王峥，陈世峰，聂海涛，段晓峰，范艳根，毛成华

挂载智能芯片的边缘视觉模组技术规范

1 范围

本标准规定了挂载智能芯片的边缘视觉模组的产品架构、系统功能要求、性能测试方法、检验标准的要求。

本标准适用于各类挂载智能芯片的边缘视觉模组。

注：在有相关的专用产品标准的情况下，产品标准优先于本标准。

2 规范性引用文件

基于 FPGA 的高能效比 LSTM 预测算法加速器的设计与实现 张奕玮
轻量化卷积神经网络技术研究 毕鹏程
基于 FPGA 的高速浮点 FFT/IFFT 处理器设计与实现 苏斌
FPGA 的模块化设计方法 张松
基于深度学习的图像去噪方法研究综述 刘 迪
基于深度学习的单图像超分辨率重建综述 邢苏霄
基于深度学习的图像超分辨率研究综述 李洪安
基于深度学习的图像超分辨率研究综述 李洪安
基于深度学习的图像语义分割方法综述 王可
基于全卷积网络的图像语义分割方法综述 李梦怡
语义分割评价指标和评价方法综述 于营

3 术语和定义

3.1

比特流 Bitstream

一个比特流是一个比特的序列。一个字节流则是一个字节的序列，一般来说一个字节是8个比特。

3.2

计算机微体系结构 Computer Microarchitecture

计算机微体系结构，也被叫做计算机组织，微架构使得指令集架构可以在处理器上被执行，指令集架构可以在不同的微架构上执行。

3.3

边缘计算 Edge computing

边缘计算，是一种分布式运算的架构，将应用程序、数据资料与服务的运算，由网络中心节点，移往网络逻辑上的边缘节点来处理。

3.4

内嵌式存储器标准规格 Embedded Multi Media Card，eMMC

eMMC (Embedded Multi Media Card）是 MMC 协会订立、主要针对手机或平板电脑等产品的内嵌式存储器标准规格。

3.5

非易失性存储设备　Flash Memory, FLASH

非易失性存储设备，是一种电子式可清除程序化只读存储器的形式，允许在操作中被多次擦或写的存储器。

3.6

噪声 Noise

噪声在图像中是指存在于图像数据中的不必要的或多余的干扰信息。噪声的存在严重影响了遥感图像的质量，因此在图像增强处理和分类处理之前，必须予以纠正。

3.7

非线性函数 Non-linear Function

因变量与自变量之间的关系不是线性的关系。

3.8

光电探测器　Photodetector

在可见光或近红外波段主要用于射线测量和探测、工业自动控制、光度计量等。

3.9

张量　Tensor

用来表示在一些向量、纯量和其他张量之间的线性关系的多线性函数。

4　缩略语

AXI	高级可扩展接口	Advanced eXtensible Interface
CIS	接触式图像传感器	Contact Image Sensor
CNN	卷积神经网络	Convolutional Neural Networks
CPU	中央处理器	Central Processing Unit
DRAM	动态随机存取记忆单元	Dynamic random-access memory
FPGA	现场可编程逻辑门阵列	Field-programmable gate array
FPU	浮点执行单元	Floating-point Processing Unit
GPU	图像处理器	Graphics Processing Unit
NLP	自然语言处理	Natural Language Processing
GRU	门控循环神经单元	Gated Recurrent Unit
I/O	输入和输出	Input/Output
ISP	图像信号处理器	Image Signal Processor
LED	发光二极管	Light-Emitting Diode Light
LSTM	长短期记忆网络	Long Short Term Memory networks
mAP	均值平均精度	mean Average Precision
MIoU	平均交并比	Mean Intersection over Union
NMS	非极大值抑制	Non-Maximum Suppression
PSNR	峰值信噪比	Peak Signal to Noise Ratio
RNN	循环神经网络	Recurrent Neural Network
SDK	软件开发工具包	Software Development Kit
SSIM	结构相似度	Structural SIMilarity
TPU	张量处理单元	Tensor Processing Unit

5　边缘视觉模组技术架构

5.1 总体技术架构

挂载智能芯片的边缘视觉模组技术架构包含硬件层、架构层、工具层及应用层，其如图1所示。

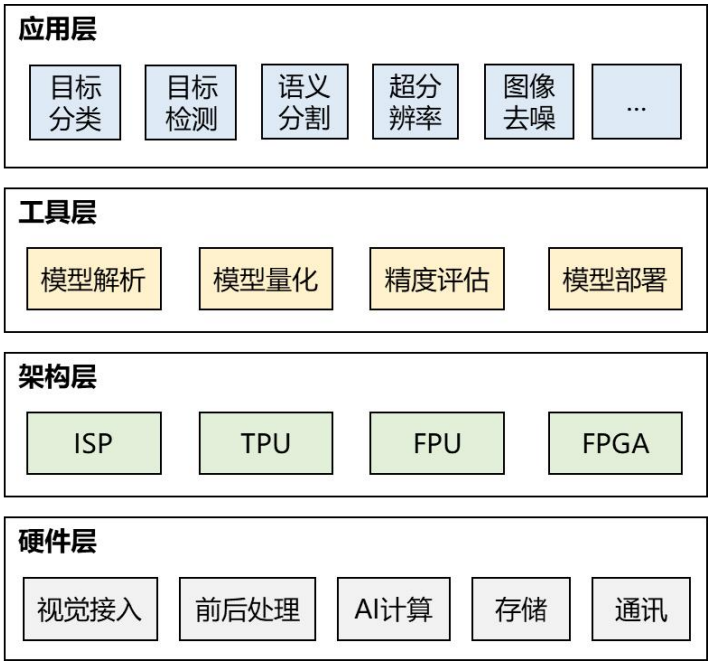

图1 挂载智能芯片的边缘视觉模组技术架构

硬件层规范底层器件结构与功能，基于光电传感器的视觉接入，进行智能化计算后与上位机通讯。架构层定义系统所需算力的部署设备，包含通用算力，专用算力及可重构算力。工具层提供算法到架构映射的主要软件及流程。应用层与工具层对接，分析场景与算法并向下层进行转换。

5.2 硬件层

硬件层由视觉接入、前后处理、AI计算、存储与通讯模块构成。

原始图像信号通过光电探测设备实现接入，经过前处理模块对图像进行优化并完成格式转换，之后图像送入AI计算模块进行智能分析，分析结果通过通讯模块传输至上位机等节点。

视觉接入模块主要包含光电成像设备，其主要由镜头及CIS图像传感器构成，其可感知可见光，红外，偏振光，雷达等影像。

前处理主要包含图像处理流水线，图像格式变换，图像尺寸向AI计算适配等。

AI计算部署于TPU及FPU处理单元，主要基于人工神经网络进行目标检测；图像的增强，如超分辨率、去噪、去斑等；图像语义分割，如像素级别的分类。

后处理主要包含张量反量化，Softmax归一化指数函数，NMS非极大值抑制等，其对AI计算的输出张量进行数学分析。

存储模块主要包含非易失存储，如FLASH,eMMC等，用来存储处理器加载的程序及经网络模型。及易失性存储，如DRAM，PSRAM，用来存储计算过程中产生的中间数据。

通讯模块主要用来向边缘模组外部发送数据，其包含原始图像及AI分析的结果，通讯可以有线方式，如以太网，USB等，也可以通过无线方式，如WIFI，蓝牙等。

5.3 架构层

架构层主要定义边缘智能视觉模组所需算力的计算机微体系结构级执行与映射方案。其主要包含四类典型处理器模型，ISP（图像信号处理器）、TPU（张量处理单元）、FPU（浮点执行单元）及FPGA（现场可编程阵列）。

ISP将光电探测器输出的Bayer格式数据转换为YUV或RGB格式，其主要包含坏点校正、黑电平校正、镜头阴影校正、RAW域降噪、白平衡增益、RGB差值、Gamma校正、颜色校正、颜色空间变换、颜色降噪等环节。

TPU对ISP输出的RGB图像进行人工神经网络处理，其主要以定点计算的形式支持算子包含张量通道调整、图像转滑窗（IMG2COL）、常规卷积及变种卷积、常规池化及变种池化、全连接、矩阵相乘、残差、张量拼接、张量量化与反量化等。

FPU为浮点运算单元，在架构层主要用来弥补TPU定点计算过程中的精度不足问题，主要用来部署非线性函数，如Exponent, Logarithm, Sigmoid, Tangent, Mish等，包含定点至浮点，浮点至定点的数据格式变换，此外可支持高精度张量计算，如循环算子（RNN, GRU, LSTM等）。

FPGA依靠其现场可重构功能，主要完成两项功能，通讯接口及前后处理。通讯接口方面，其可链接MIPI或SDI格式视觉输入，DRAM、FLASH、EEPROM等存储颗粒，以及WIFI、以太网、USB等上位机通讯接口。在前后处理方面，其可实现傅里叶变换、非极大值抑制、Softmax函数等功能。

5.4 工具层

工具层主要包含人工智能算法模型部署环节中的重要工具组件，其以软件包形式存在，主要包含模型解析、量化、精度评估、部署等组件。

模型解析工具将主流神经网络框架下的算法模型初步解析成加速器指令，以抽象的语法树形式存在。其主要包括算子参数解析、结合硬件体系结构与算子特征的硬件层算子融合；以及用于支持残差等特性的硬件层级输入输出流图编码等环节。

模型量化工具支持多种可配置的量化模式，如位宽级别的INT8/INT16量化以及权重级别的张量级别或通道级别量化。具体地，量化工具包含定点模型转换、量化参数求解、量化可行性粗评、量化参数定点化求解的量化全流程。

模型的精度评估由工具层中的仿真器实现，用于在部署/设计前的推理精度评估以及作为电路设计阶段的验证参考模型。仿真器采用与加速器相同的硬件融合层级别的执行粒度，在解析指令时调用其内部实现的参考模型库进行计算。参考模型库是在网络算子或向量操作等更高的抽象级进行建模以减少仿真时间，其具体的计算位宽和量化计算方式，应参照AI加速器微架构实现。

模型的部署工具包括转码工具，驱动工具。模型编译转码主要是依据软件SDK中的全局指令集定义文件，将指令的各个位段进行编码，同时对权重等模型数据进行二进制转码。软件驱动工具则封装了比特流烧写、模型数据流的内存预加载、硬件计算核的使能驱动、结果取回等部署全流程函数。

5.5 应用层

应用层主要面向边缘智能视觉模组使用方法，支持对各种实际场景的算法需求，其包括但不限于目标分类、目标检测、语义分割、超分辨率、去噪等具体应用。

目标分类网络是依据预定义的实例类别，判断输入图像的类别，常见的目标分类网络有AlexNet、VGG系列等。

目标检测网络是对输入图像内检测预定义类的所有实例，快速确定图像的对象和对象所属的类别，并且在检测到的实例周围绘制边界框。常见的目标检测网络有YOLO系列、Fast-RCNN、SSD等。

语义分割网络是通过提取图像的低级语义和高级语义，以像素点级的语义信息对图像的每一个像素点进行分类，确定每个像素点所属的类别，从而对图像区域进行划分。常见的语义分割网络有FCN、U-Net、DeepLab系列等。

超分辨率网络是将给定的低分辨率图像重建成具有丰富图像细节和清晰纹理的高分辨率图像的网络。常见的语义分割网络有SRCNN、FSRCNN、EDSR等。

图像去噪网络是针对图像拍摄成像过程中雾霾天气等外界噪声引起的清晰度下降，利用图像序列的上下文信息去除噪声，还原出理想情况的图像，从而提高图像清晰度。常见的图像去噪网络有FFDNet、PANet、OverNet等。

6 边缘智能模组硬件层技术要求

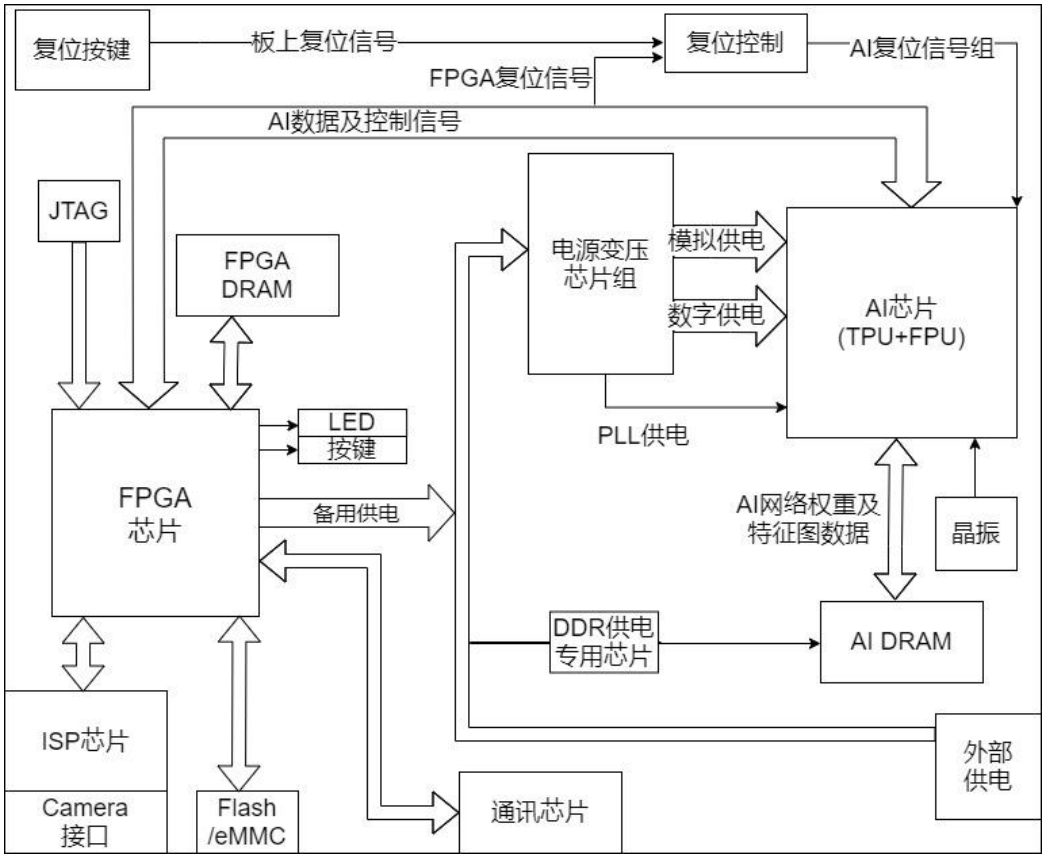

图2 边缘智能模组硬件层结构图

6.1 模块意义

硬件层为核心芯片如AI芯片，FPGA，ISP等提供工作和协同条件。总体上，分为供电部分，板上资源和辅助芯片组三个部分。供电组主要负责提供除了FPGA核心板以外的单独的电源供给。板上资源主要是按键和LED，方便查看电源状态和核心芯片的工作状态。辅助芯片组主要负责提供核心芯片工作环境。

6.2 功能要求

1） 核心芯片：

a) 计算芯片：AI芯片、ISP芯片、FPGA芯片，具体功能要求见本标准第7节“边缘智能模组架构层技术要求”。
b) 存储芯片：DRAM、Flash、eMMC等。
c) 数据接口芯片：Camera、USB、以太网、WiFi接口芯片。

2） 供电部分：要求每个电压值能够提供单独的从总电源接口处通过变压进行获取。保证及时对应某个压值的供电网络损坏，也不影响其他网络进行反馈。其中
a) 在核心供电中，要保证正确的启动顺序，可通过读取每个芯片的供电完整信号和进行使能信号控制来实现设计。
b) 要根据不同需求放置余量资源，方便后期调整供电方案。
c) 不同网络之间进行适当的过滤和隔离。
d) 设置保护资源，如防反接，防止超功率等功能需要实现

3） 板级资源：要求根据实际情况设置板上的外设资源，包括按键，LED灯，时钟资源等。
a) 通过设置按键进行复位信号来源的切换，以此来分别测试不同核心模块的工作状态。
b) 单独配置提供给FPGA按键和信号灯资源，从而使后期能够让FPGA更好的和使用者进行交互。
c) 给不同的供电网络配置供电测试灯，以此观测上电状态。
d) 配置不同信号的接口和机械模型，需符合生产工艺的参数。

4） 辅助芯片：
a) 单独配置DRAM的供电方案，根据不同型号的存储颗粒配置对应供电网络。
b) 引入晶振等时钟资源，给AI芯片提供除时钟信号。

7 边缘智能模组架构层技术要求

7.1 图像信号处理器技术要求

7.1.1 模块意义

图像信号处理器（ISP）对光学传感器输出的Bayer域数据进行信号处理，使之成为符合人眼真实生理感受的信号，并加以输出。其包含三个域的处理流程：

1) Bayer域的信号处理，主要目的是修正传感器物理特性造成的数据偏移，并且将信号进行插值，恢复完整RGB信号。
2) RGB域的信号处理，主要目的是对色彩进行补偿，恢复人眼真实感受
3) YUV域的信号处理，主要目的是分离亮度信号和色度信号，分别对两种信号处理，并且进行JPG的压缩编码。

7.1.2 功能要求

常见ISP的功能包含但不限于如下：

1) 黑电平矫正（Black Level Correction）：8bit数字信号量化范围是0-255，0代表最黑，255代表最亮。由于光学传感器要上电感光，所以最黑的情况也是有非0电平。该步矫正非0黑电平。
2) 光偏移矫正（Shading Correction）：由于镜头对不同色光的折射率是不同的，RGB三色光不能完全一致地在Sensor上成像，该步进行补偿矫正。
3) 通道矫正（Channels Correction）：在Bayer域绿色通道有两个，但是由于光学传感器的构造，感光时这两个通道会存在差异。这一步矫正此差异。

4) 白平衡（White Balance）：由于光源不同的色温会带来不同的成像，而往往人们需要去除色温带来的差异，即需要白平衡完成操作。
5) 去马赛克（Demosaic）：从不完整的颜色样本插值生成完整的颜色样本，实现Bayer排布方式转为RGB排布方式。
6) 色彩矫正（Color Correction）：利用标准色卡进行信号的矫正，使信号恢复人眼真实感受。
7) 伽马矫正（Gamma）：对信号进行非线性矫正，提高图像对比度，并使其符合显示器及人眼感受的非线性显示。
8) 降噪（Noise Reduction）：由于伽马矫正，光偏移矫正等操作会增强信号，一些噪声也随之被增强，要去除增强的影响，同时还原更干净的成像，需要进行图像降噪。
9) 边缘增强（Edge Enhancement）：因为降噪处理会平滑图片，为了让图片更为清晰，在降噪后通常需要增强图片。

7.2 张量处理器技术要求

7.2.1 模块意义

张量处理器（TPU）运行人工神经网络算法，对张量（如图像信号）进行目标识别、检测、跟踪、语义分割等智能化操作。典型TPU采用定点计算配合大量并行计算节点，其在计算精度上弱于可支持浮点运算的中央处理器（CPU）及图像处理器（GPU），但在算力、能效比（算力/功耗）、面积等指标具有一定优势。支持推理运算的TPU较适合部署于边缘计算设备。

7.2.2 功能要求

常见TPU的功能包含但不限于如下：

1) 图像转滑窗（Img2Col）：将从ISP输出的RGB数据排列格式转化为卷积操作所需要的基于滑动窗口的数据排列格式。
2) 卷积（Convolution）：通过两个函数f和g生成第三个函数的一种积分变换，表征函数f与g经过翻转和平移，实现重叠部分函数值乘积对重叠长度的积分。
3) 池化（Pooling）：又称为下采样，其压缩输入特征图，导致参数减少，简化卷积网络计算时的复杂图，同时保持了特征图特征的不变性。
4) 全连接（Full connection）：每一个结点都与上一层的所有结点相连，用来把前边提取到的特征综合起来，数学表示为向量（特征图）与矩阵（网络参数）的乘积。
5) 矩阵乘（Matrix multiplication）：两个矩阵的通用乘操作。
6) 残差（Residue）：两个特征图的点加运算，实现低层特征图直接传播到高层，其在一定程度上解决网络退化问题。
7) 拼接（Concat）：将两个及以上的特征图按照在通道等维度上进行拼接。
8) 线性整流（ReLU）：以斜坡函数及其变种为代表的非线性函数，为一种简化的神经元激活函数。
9) 上采样（Upsample）：又称图像差值，目的是放大原特征图,从而可以显示在更高分辨率的显示设备上，功能与池化相反。
10) 量化（Quantization）：将神经网络算法模型中各张量（特征图与网络参数）的浮点数表示范围简化为定点数表示范围所需要进行的数学操作。不同张量具有不同的量化参数。

7.3 浮点执行单元技术要求

7.3.1 模块意义

浮点执行单元（FPU）负责高精度的向量/张量运算，以及一些特殊函数或者非线性激活函数的近似运算。在涉及对精度有高要求的应用场景，例如语义分割，自然语言处理(NLP)等，TPU定点计算精度不足的缺陷会被放大。因此FPU的高精度浮点运算、较高的吞吐率以及支持的计算类型宽泛等特点，能够很好的弥补TPU的不足，同时与TPU协同工作在高精度和高性能的环境下。

7.3.2 功能要求

常见FPU的功能包括但不限于如下：

1) 基础的四则运算(Elementary arithmetic operation)：基础的四则运算是通用数学计算的基础，在FPU中配备了专门的浮点单元负责此类运算，如浮点加、减、乘、除等。
2) 指数运算(Exponent operation):指数运算常作为各类复杂函数的基础，如softmax,sigmoid等函数。
3) 对数运算（Natural logarithms）：以2为底的对数运算常用于各类科学计算中。
4) Sigmoid函数（sigmoid function）：Sigmoid函数由于其出色的数学特性且易于部署，被广泛应用于各类神经网络中。
5) Mish函数(Mish function): Mish函数作为最近提出的激活函数，被证实可以提高许多深度神经网络的精度结果。
6) Softmax函数(Softmax function): Softmax的数学形式中包含了大量的指数运算。常用作神经网络的最后一个激活函数，以将网络的输出归一化为概率分布。
7) 循环算子（如RNN，LSTM, GRU）: 循环算子以矩阵/向量运算为基础，与常见的张量运算不同，此类算子在应用时对精度的要求较高，常用高精度浮点运算实现。

7.4 现场可编程阵列技术要求

7.4.1 模块意义

现场可编程阵列(FPGA)依靠其数字逻辑的可重构能力实现系统全局初始化配置、对接各光电信号格式的输入接口逻辑、信号预处理逻辑、ISP/TPU/FPU信号链的中间数据缓冲、神经网络权重等系统参数导入、智能应用相关的定制后处理逻辑，以及数据发送、接收等。

7.4.2 功能要求

常见FPGA的功能包括但不限于如下：

1) 全局初始化：FPGA可通过SPI、I2C等总线引导系统各硬件模块的初始化，包含ISP，TPU，FPU等计算单元，图像传感器，DRAM，Flash，通讯模块等。
2) 信号输入：通过FPGA上的MIPI RX IP，或并口I/O接口接入图像传感器或ISP的输出数据。
3) 信号预处理：主要包含面向智能视觉应用的图像分辨率调整，特征图通道扩展，以及YUV2RGB的信号格式转换等。
4) 中间数据缓冲：通过小容量片上SRAM/BRAM高速存储器或大容量片下DRAM主存储器实现信号链中各类型中间数据的暂存，例如ISP的输出数据，预处理逻辑的输出数据，TPU与FPU的输出数据，带传输的数据等。
5) 参数导入：智能计算采用神经网络中含有大量的网络参数，其需求为断电不丢失，需要存于EPROM，Flash，eMMC等非易失存储器中。FPGA中存储接口IP可对接该类非易失存储器。

6) 定制后处理：TPU的输出数据通常为三维张量，后处理过程将其解译为使用者可理解的数据格式。例如目标检测后处理通过NMS将张量转化为一系列目标的坐标框及置信度，目标跟踪需要通过卡尔曼滤波（Kalman filter）实现目标运动趋势的预测。该类专用后处理逻辑适合于在FPGA上实现。

7) 数据发送及接收：FPGA可对接各通讯芯片，有限通讯如USB、以太网，无线通讯如WiFi及蓝牙等。处理后数据可以通过该链路发送，同时配置信息可通过该链路输入至FPGA。

8 边缘智能模组工具层技术要求

8.1 模型解析技术要求

8.1.1 框架兼容性要求

1）解析工具应具备兼容至少2种上层主流深度学习框架的能力，包括但不限于Keras/TensorFlow，Pytorch/Caffe，ONNX等；

2） 解析工具应具备兼容至少3种常见主流的工业级视觉神经网络模型的能力，包括但不限于用于分类的MobileNet系列，ResNet系列、用于检测的yolo系列、用于超分辨率的EDSR网络模型等。

8.1.2 算子兼容性要求

模型解析应具备对计算流图中算子的参数提取功能，包括但不限于上述网络中所必须的卷积类、池化类、上采样、残差、融合等标准算子集合，此外应能够支持自定义算子在原始计算流图中的嵌入。

8.1.3 计算图优化要求

模型解析工具需要对所兼容框架下的网络模型进行进一步的计算图优化，应具备如下要求：

1）解析工具应具备对原始模型计算图进行优化的能力，具体包括但不限于算子融合、算子结果内存原位共享等；

2）解析工具应针对硬件特性对原始模型中的算子进行数据依赖分析功能；

3）解析工具应能够合理地将原始计算图混合映射到架构层的多个硬件计算核，即根据精度、速度要求分发到模组架构层的TPU和FPU等。

8.2 模型量化技术要求

8.2.1 模型量化方案要求

在资源受限的边缘侧，模型量化工具应采用合理的量化方案来减少内存访问并减少硬件设计开销，具体要求如下：

1）应采用离线量化（训练后量化）的量化级别，提前确定好量化参数，在不修改训练模型的前提下完成量化；

2）应采用8bit和16bit为主的量化位宽，对不同的精度场景采用对应的位宽方案或是混合精度量化。特别地，对于8bit权重量化，应采用per-channel的量化方案来弥补量化损失。

3）对于其他的量化细节，我们应对称量化方式、线性量化、确定性量化为主的方案，确保硬件侧的资源开销较少；

8.2.2 模型量化训练和部署要求

1）模型量化训练的数据集应采用至少200张以上的图片，建议采用模型训练时的验证集中的子集作为具有代表性的量化训练数据集；

2）在完成量化训练后，应首先在CPU上直接进行推理完成模型的预评估，如INT8的量化位宽，其精度相对浮点网络下降应小于5%；

3）量化工具应能够对量化训练后的浮点参数进行定点化的近似计算，其中浮点参数近似计算的误差应小于5%，近似后的定点参数位宽应不超过8bit；

8.3 模型精度评估技术要求

8.1.1 软件参考模型建模要求

软件参考模型用于对架构层中以TPU为主的各种计算核进行建模，应符合如下要求：

1）参考模型的输入激励包括模型数据、输入数据、指令数据在内应与架构层保持完全一致；

2）参考模型的顺序应采用硬件融合层级的执行粒度，层级执行顺序应保持与架构层完全一致，

3）参考模型的总体建模级别应处于寄存器传输级别之上的系统级。其中，对计算库中的各个硬件层融合层，以算子进行建模；对各个算子，以软件向量和标量操作进行计算；

8.1.2 仿真器指标要求

1） 精度方面，仿真器最终推理结果以及融合层级的中间输出输出结果应尽可能与架构层硬件保证完全一致，最大误差应小于1%；

2） 性能方面，仿真器的执行算力与源主机平台存在一定的相关性，但运行上述的网络均需要保证在10s以内完成神经网络的仿真。

8.1.3 仿真器功能要求

1）仿真器应支持算子级别的中间结果的取回与分析比较；同时支持每一层的精度仿真结果的误差可视化；

2）仿真器中的算子实现，除常规算子计算操作，还应该实现与定点TPU中一致的量化定点恢复、截断饱和等量化计算操作；

3）仿真器应在仿真程序保证足够的灵活性，具体要求具备为执行计算图中的任意中间子图或者任意层的仿真推理的能力；

8.4 模型部署技术要求

8.4.1 部署层转码要求

1）部署层应支持对计算图在架构层中映射后的指令进行文本生成和编码，其保存结果格式为二进制字符串的txt文件。

2）部署层应支持将模型网络数据进行编码，其保存结果格式为二进制字符串txt文件。

8.4.2 部署层硬件接口要求

1）部署层应可提供接口驱动硬件完成内存数据加载、计算张量结果取回等主机数据流通讯操作等功能；

2）部署层应可驱动硬件固件接口层进行配置，在此基础上封装包括多种引擎核推理操作函数，包括复位以及控制与状态寄存器使能和读回等功能；

3）部署层应可驱动模组中可重构阵列（FPGA）完成比特流的烧写和配置。

8.4.3 部署层应用接口要求

1）应支持主流视觉程序接口如OpenCV等以满足输入输出数据处理要求；

2）应向应用层提供模型加载、硬件推理驱动、结果分析等高层次API函数接口；

3）结果对比分析应包含至少如下两个层次，一是原始模型浮点结果与边缘加速定点结果反量化后的向量相似性分析对比；边缘推理结果与实际应用级别的推理结果对比（详见第九章算法层要求）

9 边缘智能模组算法层技术要求

9.1 目标分类网络技术要求

Top-1错误率即对一个图片，如果概率最大的是正确答案，才认为正确。Top-5错误率即对一个图片，如果概率前五中包含正确答案，即认为正确。

由于在边缘设备部署目标分类网络采用不同种类的数据集训练、实例种类、网络模型量化或网络模型压缩等因素导致实际效果与理论效果存在较大的偏差。因此，语义分割网络使用测试集ImageNet，实现的理论效果作为参考，应满足如下技术要求：

——Top-1错误率理论效果一般不低于60%

——Top-5错误率理论效果一般不低于80%

9.2 目标识别网络技术要求

目标检测网络是对输入图像内检测预定义类的所有实例,快速确定图像的对象和对象所属的类别，并且检测出实例在图像上的位置和大小，检测到的实例周围绘制边界框。目标识别网络主要以均值平均精度（mean Average Precision，mAP）去衡量目标检测网络实际的效果。

均值平均精度是目标检测算法的主要评估指标。目标检测模型通常会用精度(mAP)指标描述优劣，mAP值越高，表明该目标检测模型在给定的数据集上的检测效果越好。

由于在边缘设备部署目标分类网络采用不同种类的数据集训练、实例种类、网络模型量化或网络模型压缩等因素导致实际效果与理论效果存在较大的偏差。因此，语义分割网络使用测试集PASCAL VOC 2012，实现的理论效果作为参考，应满足如下技术要求：

——mAP理论效果一般不低于60%

9.3 语义分割网络技术要求

语义分割网络是通过提取图像的低级语义和高级语义，以像素点级的语义信息对图像的每一个像素点进行分类，确定每个像素点所属的类别，从而对图像区域进行划分。语义分割从微观的角度可以理解为将图像中的每一个像素点进行分类，是一种像素级的空间密集型预测任务。语义分割试图在语义上理解图像中每个像素的所代表的含义；从宏观的角度则可以将语义分割看作是将一致的语义标签分配给一类事物，而不是每个像素。

平均交并比（Mean Intersection over Union, MIoU）是当前语义分割研究中最常用的指标。在语义分割中，交并比表示预测掩码与标签像素的交叠率，评价预测的前景区域是否精准。平均交并比是计算每个类别中交互比值的算术平均值，用于总体数据集的像素重叠情况。相较其他的语义分割评价指标而言，MIoU在不同数据集的得分更稳定，是更标准的分割评价指标。

由于在边缘设备部署语义分割网络采用不同种类的数据集训练、使用场景、实例种类、网络模型量化或网络模型压缩等因素导致实际效果与理论效果存在较大的偏差，因此，语义分割网络使用测试集CamVid，实例种类为32时，实现的理论效果作为参考，应满足如下技术要求：

——MIoU理论效果一般不低于45%

9.4 超分辨率网络技术要求

超分辨率网络是能够将输入的低分辨率图像重建成具有丰富图像细节和清晰纹理的高分辨率图像的网络，不仅可以提高图像分辨率改善图像质量，还可以辅助边缘设备完成其他的机器视觉任务，有助于提升目标检测、图像去噪等其他神经网络的性能。

本标准的超分辨率神经网络均采用全参考型的客观评价指标来评估超分辨率网络实现的效果，客观评价指标是指通过不同的数学模型和算法来评估图像质量的方法，具有简单、高效、可重复性强等优点。全参考型的客观评价指标是将重建的高分辨率图像与真实的高分辨率图像进行比较计算得出的，该评价指标包括峰值信噪比(Peak Signal-to-Noise Ratio，PSNR)、结构相似度(Structural SIMilarity，SSIM）。

PSNR 是信号的最大功率和信号噪声功率之比，用来测量已经被压缩的重构图像的质量，通常以分贝(dB)来表示。SSIM 是一种用来衡量两幅图像之间的相似度的指标，其取值范围为[0，1]。PSNR 和 SSIM 的值越大，表示重建图像的质量越好，即超分方法的性能越好。

由于在边缘设备部署超分辨率网络采用不同种类的数据集训练、缩放因子不同、输入的低分辨率图像存在较大的差异、网络模型量化或模型压缩等因素导致实际效果与理论效果存在较大的偏差，因此，超分辨率网络使用测试集SET5，缩放因子为2时，实现的理论效果作为参考，应满足如下技术要求：

——PSNR理论效果一般不低于33dB

——SSIM理论效果一般不低于92%

9.5 图像去噪网络技术要求

图像去噪网络是针对图像拍摄成像过程中雾霾天气等外界噪声引起的清晰度下降，利用图像序列的上下文信息去除噪声，还原出理想情况的图像，从而提高图像清晰度。部署边缘设备的环境中的真实噪声是指由拍照设备在照明条件差、相机抖动、物体运动、空间像素不对准、颜色亮度不匹配等情况下获取的图像中存在的噪声，并且具有噪声水平未知、噪声类型多样、噪声分布复杂且难以参数化等特点。

由于在边缘设备部署图像去噪网络采用不同种类的数据集训练、噪声水平不同、输入的低分辨率图像存在较大的差异、网络模型量化或模型压缩等因素导致实际效果与理论效果存在较大的偏差，因此，图像去噪网络使用测试集SET12，噪声水平为35时，实现的理论效果作为参考，应满足如下技术要求：

——PSNR理论效果一般不低于25dB

——SSIM理论效果一般不低于0.70