基于自主 AI 芯片的多算法融合应用全流程处理效率优化及实战应用

■ 文 / 刘祎璠[1] 王生进[1] 赵军[2] 张来清[3] 杨旭[4]
1. 清华大学媒体大数据认知计算研究中心 2. 太原市公安局
3. 永州市公安局 4. 北京欣博电子科技有限公司

摘 要： 多算法融合应用是指在单个计算单元上，同时运行人脸检测、人脸识别、目标检测、行人再识别（ReID）等算法模型，实现多重技术手段的跨视域连续跟踪和识别。由于多算法融合应用的技术难度和算力要求均较高，使得现有的多算法模型融合应用技术大多使用多服务器组合完成，少有多算法融合模型应用在单卡上的落地应用。同时，以往对加密视频流的分析一般需要经过解密、解码与分析分三步在不同设备上进行，使得对加密视频分析整体效率远远低于普通视频流。笔者基于国内自主研发的具有自主知识产权的 Anrui-810 芯片的边缘端计算模组 Anrui-C1 和云端计算卡 Anrui-P20，对多算法融合应用及加密视频流分析效率进行了优化并落地试用，在永州市雪亮工程项目中及太原市试点的实际监控场景中进行了实战应用。实战结果显示，自主 AI 芯片方案对比海思 Hi3559A、英伟达 Jetson Xavier NX 与英伟达 T4 方案在全流程完整应用时占有大幅优势，尤其对加密视频流分析效率优势更加突出。

关键字： 自主 AI 芯片 多算法融合 加密视频分析

1 引言

多算法融合应用集成了视觉AI技术中各种先进算法，可以从多维度对视觉目标进行判断，在公共安全和智慧安防领域有着广泛的应用前景。该任务是当前计算机视觉领域十分具有挑战性的研究课题，也是在公共安全和智慧安防领域一个具有重要需求和广泛应用前景的技术。由于多算法模型融合应用技术难度和算力要求较高，使得现有的多算法模型融合应用技术大多使用多服务器组合完成，鲜有相关的多算法融合模型应用在单卡上落地应用。在实际应用中，大量场景不具备部署多服务器的物理空间条件，或者部署成本太高，使得多算法模型融合应用并不常见。近年来，随着计算能力的增强，基于深度学习的 AI 算法的性能远超传统方法并成为了目前市场应用的主流。Sun et al. (2018) 提出了基于不同行人部件的行人再识别算法，Zhang et al. (2021) 提出了基于时空 Transformer 的多帧行人信息的融合方法，Zhou et al. (2020) 提出了基于 GAN 网络和对抗学习的跨域行人再识别方法，Gu et al. （2020）提出了基于帧间部件特征对齐和 3D 卷积网络的多帧行人再识别算法。

传统视频监控视频流都是裸传，非常容易发生视频数据泄露等问题。因数据丢失造成的公民隐私泄露、政府关键信息被窃取事件时有发生。通过增加防火墙、边界等网络安全手段难以完全防范数据丢失的发生，为此，我国出台了《公共安全视频监控联网信息安全技术要求》（国标代号 GB 35114-2017，以下简称 GB 35114）强制性国标，首次以数据签名与加密的手段来保护数据的安全。目前，行业内对符合 GB 35114 标准加密的视频流进行分析的解决方案较少，一般需要经过解密、解码与分析这

三个步骤在不同设备上进行，使得对加密视频的分析整体效率远远低于普通视频流。在大量 GB 35114 强制性国标普及应用的场景下，如何对符合标准规定的 C 级加密视频流进行分析成了安防行业迫切需要解决的问题。

边缘 AI 计算模组和云端 AI 计算卡是基于深度神经网络的机器学习算法落地应用的重要工具，笔者迫切需要找到一套能够进行多算法融合计算并支持国标加密视频流分析的边缘端和云端硬件解决方案。本文深入评测了目前行业内所使用的主流边缘端计算模组海思 Hi3559A、英伟达 Jetson Xavier NX 和云端计算卡英伟达 T4，以及欣博电子新发布的 Anrui-810 边缘端计算模组 Anrui-C1 和云端计算卡 Anrui-P20，并对多算法融合计算全业务流程的性能进行了优化设计，特别是对国标加密视频流的全业务流程处理过程进行了优化，并在永州市雪亮工程和太原市试点的实际监控场景中实现了加密 SVAC 视频流与非加密 H.264/H.265 视频流并行计算、相辅相成的一套兼容性系统。

2 国产自主 AI 芯片 Anrui-810

2.1AI 芯片 Anrui-810 的特点

Anrui-810 是国内自主研发并于 2020 年底正式发布的一款面向端和云的人工智能处理器芯片。该芯片开发过程严格遵循相关国家及行业标准要求，以“安全”和“智能”为基线， 采用台积电 28 纳米制造工艺和封装级多通道片上内存解决方案打造而成，其结构框图如图 1 所示。

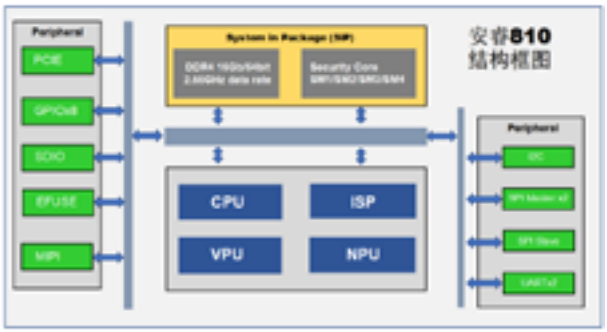

图 1 Anrui-810 系统结构框图

2.1.1 高效、灵活的神经网络处理单元

自研神经网络处理单元，具备灵活可编程能力，能效比高。全面支持各类主流算法框架，包括 caffe、pytorch、mxnet、onnx、darknet、keras、tensorflow(mlir) 等。

2.1.2 多视频格式支持

自研视频编解码处理单元，实现 SVAC2.0/H.264/H.265/JPEG 视频、图片硬件编解码，具备强大的高并发视频处理能力。

2.1.3 国密算法支持

采用 SiP 技术，封装安全加密处理器，硬件级别 SM2、SM3、SM4 算法加解密，并通过相关安全认证，满足视频图片及其他各类数据的实时加密、解密需求。

2.1.4 扩展性好

集成高速 PCIE 接口，根据实际需求可实现线性处理性能扩展。采用 SiP 技术，封装 16Gb/64bit DDR4 KGD，减少 PCB 板设计尺寸，提高板卡计算密度。

2.1.5 易用性强

内置高性能 8 核心 1GHz 主频的 RISCV 处理器，支持包括 linux 在内主流操作系统，具备完整易用的软件开发生态，方便用户进行软件应用移植。

2.2 边缘计算模组 Anrui-C1

Anrui-C1 边缘计算模组集成 Anrui-810 芯片，支持主流的深度学习推理算法，具备高性能、高可靠性等优点，可广泛适用于各种人脸检测与识别、视频结构化、视频转码处理等边缘计算场景。

C1 尺寸仅有 5x5mm，内存 2GB，支持 8~16 路 H.264/H.265/SVAC2.0 编解码与转码，6W 功耗下提供 10TOPS INT8 算力（非稀疏），其结构框图如图 2 所示。

图 2 边缘计算模组 Anrui-C1 示意图

2.3 云端 AI 计算卡 Anrui-P20

Anrui-P20 云端计算卡内置多颗 Anrui-810 芯片，是面向专业安防领域数据中心级别的高性能智能板卡产品，支持主流的深度学习推理算法，具备高可靠性、高扩展性特点，可广泛适用于各种人脸检测与识别、视频结构化等高性能计算场景。

图 3 P20 AI 计算卡示意图

P20 卡采用标准半高单槽 PCIE 设计。得益于芯片封装级内存设计方案，板卡内存有效带宽 170GByte/s，容量 16GB，单卡支持 64-128 路 H.264/H.265/SVAC2.0 全高清视频解码，50W 功耗下提供 80TOPS INT8 算力（非稀疏）。

3 多算法嵌入芯片流程

3.1 视频流加密与解密

GB 35114 标准规范了前端设备的安全能力，由弱到强分别是 A 级、B 级、C 级。C 级要求设备具备对国产视频编解码标准 SVAC 视频数据签名与加密能力。边缘端 AI 计算模组为满足 GB 35114 标准要求，需要在前端对视频数据加密；云端 AI 计算模组接收到 GB 35114-C 级视频流进行分析，需要依次进行解密、解码，产生原始视频数据进行 AI 分析。以 Hi3559A 为主芯片的边缘计算模组、英伟达 Jetson Xavier NX 边缘计算模组和英伟达 T4 云端计算卡加解密需要与其他产品配合进行，国产边缘计算模组 C1 和云端计算卡 P20 自身集成加解密和 SVAC2.0 编解码功能，提供对 GB 35114 标准支持的 SDK，可将多步运算集中在单个设备内完成。

3.2 视频分析全流程介绍

基于深度学习的人脸检测、人脸识别、目标检测、行人再识别（ReID）等算法是目前安防领域的常用算法。上述深度学习算法一般包括几个步骤：检测、特征提取和特征比对。检测即在实际的监控视频中，采用目标检测的算法框选出目标所在区域；特征提取即采用深度神经网络的方法提取图片中的语义信息，以获取有鉴别力的目标特征；特征比对即将 query 图片提取出的特征与 gallery 中的每一张图片提取出的特征进行比对，计算其相似度，并按照相似度进行排序。有效的算法应使得相同 id 的图片特征相似度尽可能高，不同 id 的图片特征相似度尽可能低。

实际操作中，本文使用 Retinaface 作为人脸检测网络，使用 Yolov4 Tiny 作为目标检测网络，使用在海量数据集 MARS 上训练好的 MobileNet、VGG16、Resnet50 网络作为特征提取器，其中 Yolov4 Tiny 网络和特征提取网络配合可应用于行人再识别（ReID）算法。

对于特征提取网络训练，记一张输入图片为 x_i，对应的类别标签为 y_i，特征提取网络为 ϕ。提取出该图片的特征：$f_i=\phi(x_i)$，再将 f_i 通过一个分类器得到预测的分类标签 $\bar{y}_i$。为了提高特征的鉴别性，本文采用 CrossEntropy Loss 和 Triplet Loss 作为损失函数对特征提取网络进行训练。CrossEntropy Loss 和 Triplet Loss 损失函数的公式如下：

$$L_{ce}=E_{(x,y)\sim(X,Y)}\left[\sum_{k=1}^{K}\bar{y}_k\log(\bar{y}_k)\right] \quad (1)$$

$$L_{tri}=E_{(x,y)\sim(X,Y)}\left[\max(0,m+\|f-f_p\|-\|f-f_n\|)\right] \quad (2)$$

其中，K 为类别数，$\bar{y}_k$ 为 0 或 1，表示这个样本是否属于第 k 类。m 为一个正的常数，f_p 和 f_n 分别表示一个与当前样本同类样本和一个不同类样本提取出的特征。CrossEntropy Loss 是一种分类损失，优化目标是使得网络提出的特征能被正确分类。Triplet Loss 是一种度量损失，优化目标是使得相同类别的样本提取出的特征间的距离尽可能小，不同类别的样本提取出的特征间的距离尽可能大。

在深度学习算法以外，实现完整的视频分析，一般还需要有视频解码、颜色空间转换与缩放过程，如果是加密视频流，还需要在解码前先进行解密。故完整的视频类 AI 应用一般处理流程如图 4 所示。

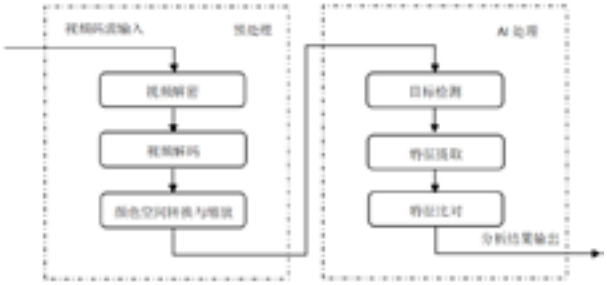

图 4 视频类 AI 应用的一般处理流程

3.3 AI 算法移植

以 Hi3559A 为主芯片的边缘计算模组、英伟达 Jetson Xavier NX 边缘计算模组和英伟达 T4 云端计算卡是主流的 AI 计算硬件，其使用方法已被业内广泛掌握。本文我们特别研究了支持 GB 35114 国标的国产边缘计算模组 Anrui-C1 和云端计算卡 Anrui-P20 的深度学习算法移植方法。因 C1 和 P20 都是基于同一颗芯片，故移植方法一样。

首先，在 pytorch 框架下完成了人脸识别、目标检测等多个算法模型的训练，得到 .pth 格式的模型参数文件。为了将 pytorch 框架下的模型嵌入到芯片中，需要对模型进行一些转换。具体流程如图 5 所示。

1）将 pytorch 框架下的 .pth 格式的模型文件转化为 .onnx 开放式模型。

2) 使用 stkn_mapper 将 .onnx 格式模型转化为 stkn 芯片支持的格式(.param, .bin）。

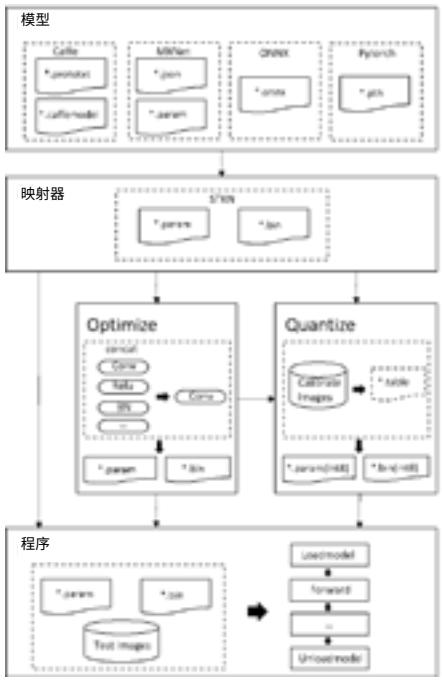

图 5 STKN 芯片移植流程图示

3) 使用 stkn_optimize 进行一些层的合并 (如 batchnorm 合并到之前的 convolution) 以减少硬件耗时。

4) 使用 stkn_quantize 进行模型的量化，先输入需要量化的 stk 模型权重和数据集，生成所需的 table 文件，再使用这个 table 文件将 fp32 的 stkn 模型权重转为 int8。

5) 完成移植，在芯片上进行测试。

4 实验结果对比及分析

4.1 实验平台

本文分别进行了边缘计算模组和云端计算卡的性能对比。边缘计算模组可以直接对比测试，云端计算卡需要搭配服务器进行测试。采用 DELL PowerEdge R940 服务器，配置如表 1 所示，分别搭配单张 Anrui-P20 加速卡和 Nvidia-T4 加速卡进行测试。为了准确反映出不同算法的真实落地性能，本实验测试项将涵盖图 4 所示完整处理流程，包括视频解密、视频解码、颜色空间转换、图像缩放、AI 处理五个环节。

表 1 实验平台配置

	边缘端 AI 测试环境			云端 AI 测试环境	
	环境 1	环境 2	环境 3	环境 4	环境 5
测试机操作系统	----			CentOS 7	
测试机处理器	----			Intel Xeon Gold 6136*4	
测试机内存	----			768GB DDR4 2933Mhz	
被测设备型号	Hi3559A	Nvidia-NX	Anrui-C1	Nvidia-T4	Anrui-P20
被测设备算力	4Tops	21Tops	10Tops	130Tops	80Tops
被测设备功耗	5w	15w	5w	70w	50w
被测设备内存	4GB 64bit DDR4	8GB 128bit LPDDR4x	2GB 64bit DDR4	16GB 256bit GDDR6	16GB 512bit DDR4

4.2 视频编解码性能

实验测试了 H.264/H.265/SVAC2.0 三种视频格式，测试序列分辨率为 1920x1080，帧率为 30 帧 / 秒。测试结果如表 2 所示。

表 2 视频解码性能

	边缘计算模组			云端计算卡	
视频格式	Hi3559A	Nvidia-NX	Anrui-C1	Nvidia-T4	Anrui-P20
H.264	16 路	22 路	16 路	30 路	128 路
H.265	16 路	44 路	16 路	30 路	128 路
SVAC2.0	N/A	N/A	16 路	N/A	128 路

从测试结果可以看出，不同格式下 Anrui-C1 视频编

解码能力与 Hi3559A 相当，Nvidia-NX 性能虽然较强，但功耗为其他两款边缘设备几倍，同等功耗下性能反而偏弱，Anrui-P20 的视频编解码性能远强于 Nvidia-T4，并且 Anrui-C1 与 Anrui-P20 可以支持我国自主知识产权的 SVAC2.0 视频码流的编解码。

4.3 综合性能

本文对多算法融合应用进行了测试，不同应用场景下的算法所用模型不同。对人脸检测，选择了 Retinaface 网络模型；对目标检测选择了 YoloV4 Tiny 网络模型；对特征提取，选择了 MobileNet、VGG16、Resnet50 网络模型。不同的网络模型有着不同的计算量和访存需求，不同硬件上的 AI 分析实际性能表现并非由算力这个单一指标决定，另外一个重要因素是计算内核的有效访存带宽。本文对边缘计算模组和云端计算卡分别进行了单跑 CNN 和全流程计算的对比测试。所用网络模型数据精度均为 INT8，视频流分辨率为 1920x1080，帧率为 30 帧 / 秒。测试结果见表 3。

表 3 a 边缘计算模组综合测试结果

应用场景	网络模型	Hi3559A	Nvidia-NX	Anrui-C1
人脸检测	Retinaface	271	759	776
目标检测	YoloV4_Tiny	160	419	402
特征提取	MobileNet_v1	373	2477	1067
	MobileNet_v2	245	1443	700
	VGG16	28.6	241	132
	Renet50	95.2	741	231
ReID	Resnet50	63.8	497	163

表 3 b 云端计算卡综合测试结果

应用场景	网络模型	英伟达 T4			安睿 P20		
		CNN(fps)	解码 + 变换 + 缩放 +CNN	解密 + 解码 + 变换 + 缩放 +CNN	CNN(fps)	解码 + 变换 + 缩放 +CNN	解密 + 解码 + 变换 + 缩放 +CNN
人脸检测	Retinaface	1937	25 路	12 路	6168	96 路	96 路
目标检测	YoloV4 Tiny	2416	25 路	12 路	3216	96 路	96 路
特征提取	MobileNet	8748	25 路	12 路	8536	96 路	96 路
	VGG16	1750	25 路	12 路	1056	30 路	30 路
	Renet50	4120	25 路	12 路	1848	60 路	60 路
ReID	Resnet50	2762	25 路	12 路	1304	43 路	43 路

注：包含解密测试项所采用的视频输入码流为 SVAC2.0，不含解密测试项输入码流为 H.265。ReID 测试时使用 Market1501 数据集。

4.4 结果分析

由表 3 a 边缘计算模组对比结果可见，Anrui-C1 相比 Hi3559 在性能跟能效两方面都表现出绝对优势，相比 Nvidia-NX 在性能方面虽有落后，但能效方面同样表现出绝对优势。结合各模组视频解码能力综合来看，同等功耗下 Anrui-C1 在三款边缘计算模组产品中胜出，这对于需要大规模部署的边缘计算应用来说具备明显的经济效益。并且由于 Anrui-C1 支持 SVAC2.0 编解码的特殊特性，使其在有 GB 35114-C 级安全要求的边缘设备计算场景中有着更广泛的应用空间。

由表 3 b 云端计算卡对比结果看，Anrui-P20 与英伟达 T4 AI 推理性能各有优势。对检测类算法运行速度 P20 明显高于 T4，对分类算法运行速度慢于 T4。参考两款产品文档，本文分析认为主要原因是 P20 采用 DDR4 内存，T4 采用 GDDR6 内存。内存读写速度差异应该是以上测试结果的主要原因。但在实际应用中，检测算法运行频次远高于分类算法，故 P20 特点更为符合实际需求。若是可以基于特定的分类网络（MobileNet）训练模型以完成相应的推理任务，例如在行人再识别（ReID）应用中部署 MobileNet 模型作为特征提取器，则 P20 综合优势会更加突出。

对 H.264/H.265 等非加密视频流进行完整流程解析，即依次进行解码、缩放、色域变换、检测、提取特征、分类或识别，总体运行速度 Anrui-P20 显著高于英伟达 T4。本文分析认为主要原因是 P20 解码能力 128 路远高于 T4 的 30 路。当解码与 CNN 同时运行时，对内存使用存在竞争，故 CNN 和解码性能均有一定程度下降。下降后 P20 整体性能可保持在 T4 性能的 4 倍左右。同时，P20 支持卡内进行色域变换和缩放，进一步提升其计算优势。

对加密 SVAC2.0 视频流进行完整流程解析，即依次进行解密、解码、缩放、色域变换、检测、提取特征、分类或识别，总体运行速度 Anrui-P20 优势相较于英伟达 T4 更加明显。本文分析认为主要原因是英伟达 T4 不支持解密和 SVAC2.0 解码，需要把视频流送至加密机 / 加密卡进行解密，然后在 Intel CPU 上进行软解码，再将 YUV 数据送至 T4 分析。软解码占用大量 CPU 资源，YUV 通

过 PCIE 传输也存在性能瓶颈。因此，在测试主机上只能达到 12 路左右性能。而 P20 完全在卡内实现，性能与处理非加密 H.264/H.265 视频流相比无任何下降。

5 实战应用

通过在永州市和太原市一类前端点位部署搭载 Anrui-C1 模组的安全接入盒，实现了对普通摄像机的 GB 35114 -C 级安全升级改造和智能化升级改造，在二、三类点社会单位部署符合《公共安全社会视频资源安全联网设备技术要求》（GA/T 1781-2021）的安全联网设备，并在公安局机房视频专网内增加了搭载 Anrui-P20 的云端计算卡的网关服务器，实现对原有 GB/T 28181 平台向 GB 35114 平台的升级改造，在视图库硬件设备中增加了搭载 Anrui-P20 云端计算卡的服务器使其增加了对 GB 35114-C 级码流的支持。示意图如图 6 所示。

图 6 “安全”+“智能”升级改造示意图

经过升级改造的系统，不仅一类点达到了 GB 35114-C 级的数据安全要求，而且低成本的增加了大量二、三类点的社会视频资源，同时增强了后端系统对视频流解密、解码和分析处理性能和兼容性。

6 结语

基于自主芯片的边缘计算模组 Anrui-C1 和云端计算卡 Anrui-P20 凭借超强的视频编解码能力和对加解密的支持，在多算法融合应用全业务流程处理场景下，综合性能对比主流产品分别有 2 倍和 4 倍的性能优势，尤其是其对 GB 35114 的支持，可以达到使现有系统实现“智能”+“安全”双重升级的效果。

参考文献

[1] L. Zheng, Zhi Bie, Y. Sun, Jingdong Wang, Chi Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identifification. In ECCV, 2016.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770－778, 2016.

[3] Alexey Bochkovskiy, Chien-Yao Wang, Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection.

[4] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, Shengjin Wang. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In ECCV, 2018.

[5] Tianyu Zhang, Longhui Wei, Lingxi Xie, Zijie Zhuang, Yongfei Zhang, Bo Li, Qi Tian. Spatiotemporal Transformer for Video-based Person Re-identification. In CVPR, 2021.

[6] Yang Zou, Xiaodong Yang, Zhiding Yu, B.V.K. Vijaya Kumar, Jan Kautz. Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification. In ECCV, 2020.

[7] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-Preserving 3D Convolution for Video-based Person Re-identification. In ECCV, 2020.

[8] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 815-823, doi: 10.1109/CVPR.2015.7298682.

[9] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In Computer Vision, IEEE International Conference, 2015.