MODEL DEVELOPMENT FOR EDGE AI

CVPR 2024 Tutorial

The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024

Seattle, WA, USA

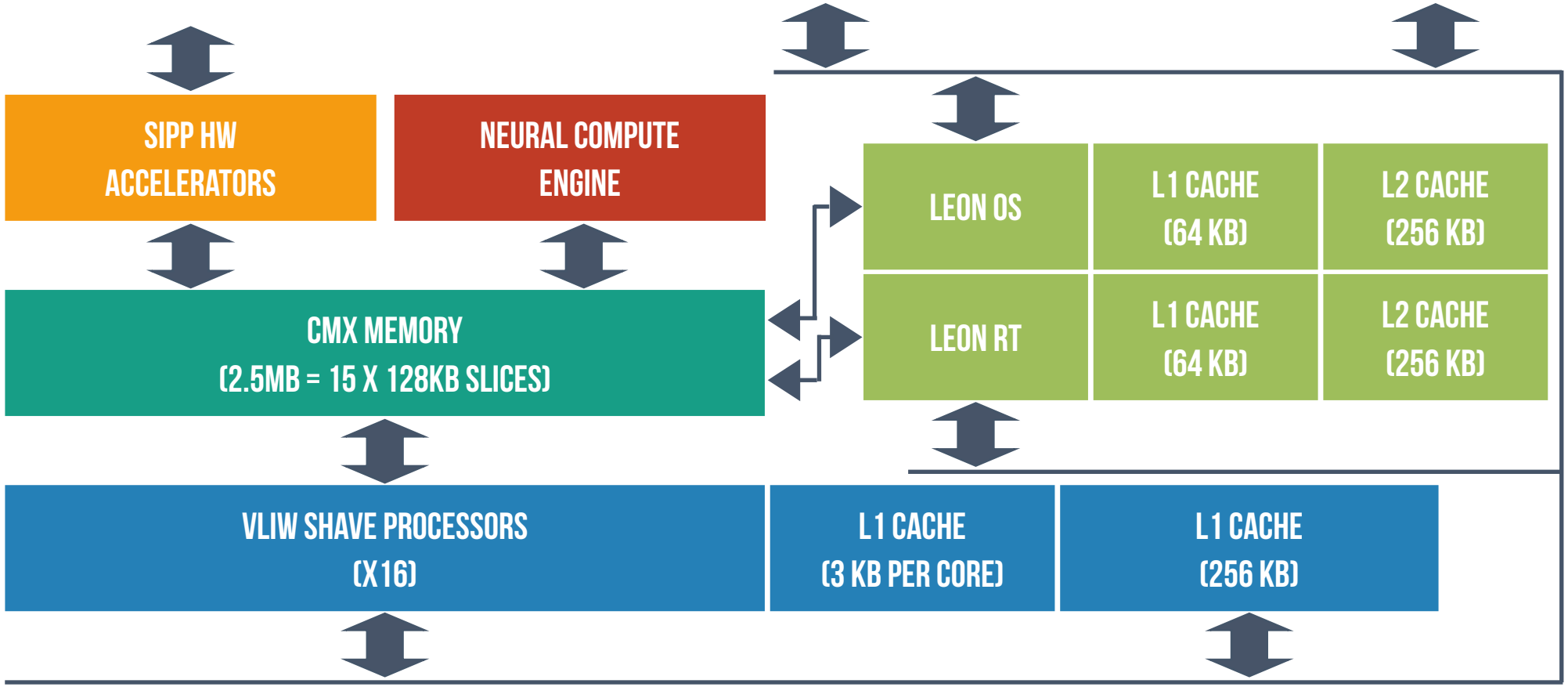
DESIGN A SEGMENTATION MODEL

Overview

1. *Hardware,* the model will be executed in a camera with Intel Movidius Myriad X VPU.
2. *Operating specifications,* the basic requirements to execute the model in an edge device in realtime.
3. *Model design,* the model architecture design considering the Edge device limitations.
4. *Dataset,* the synthetic and real data used to train the machine learning model.
5. *Training,* the process of teaching a machine learning model to make predictions or decisions.
6. *Results,* comparing our results with the models available in unified communication platforms.

INTEL MOVIDIUS MYRIAD X VPU HARDWARE

Petrongonas et al. (2021), ParalOS: A Scheduling & Memory Management Framework for Heterogeneous VPUs



OPERATING SPECIFICATIONS

Requirements

Video and audio processing pipeline and other DL models.

The body segmentation model must run on 1-2 SHAVES.

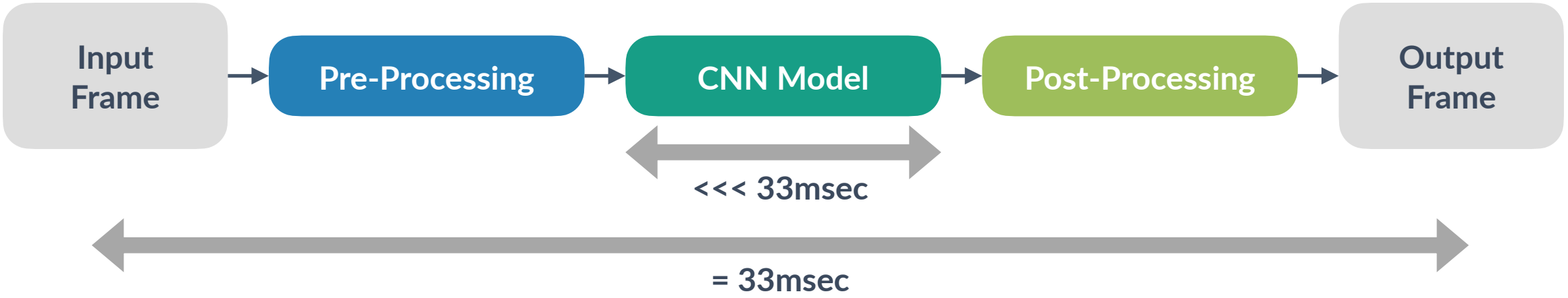
The model will segment only a single individual in the frame.

Model receives ROI crop from using the head-body detection model.

Whole segmentation pipeline must run in real-time (33msec).

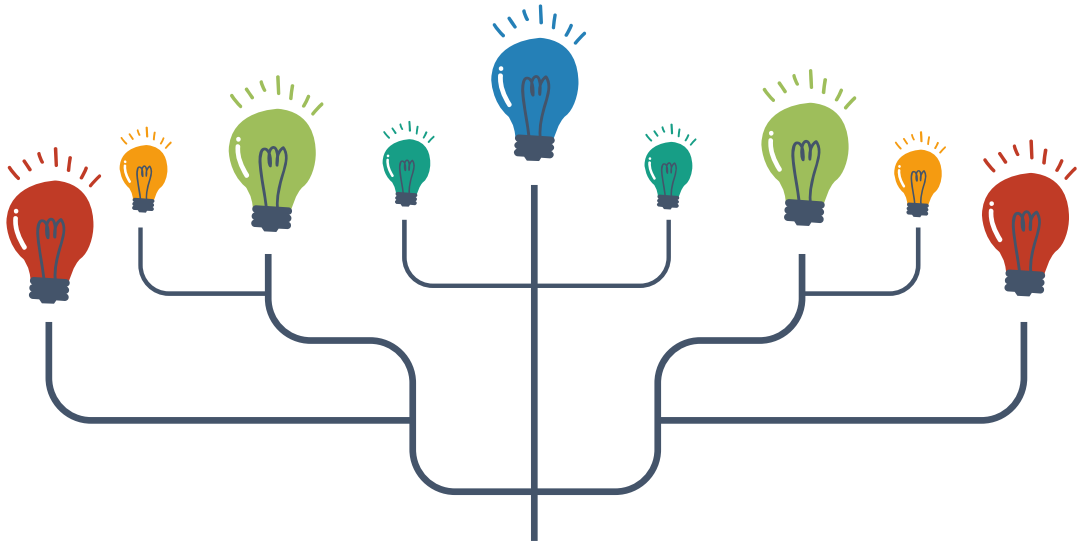
OPERATING SPECIFICATIONS

Requirements



MODEL DESIGNING

Understanding Hardware



WHAT WE KNOW FROM DOCUMENTATION

1. **Data Type**
   Only support 16bit Floating Point.

2. **NCE**
   Limited number of operations directly supported by Neural Compute Engine.

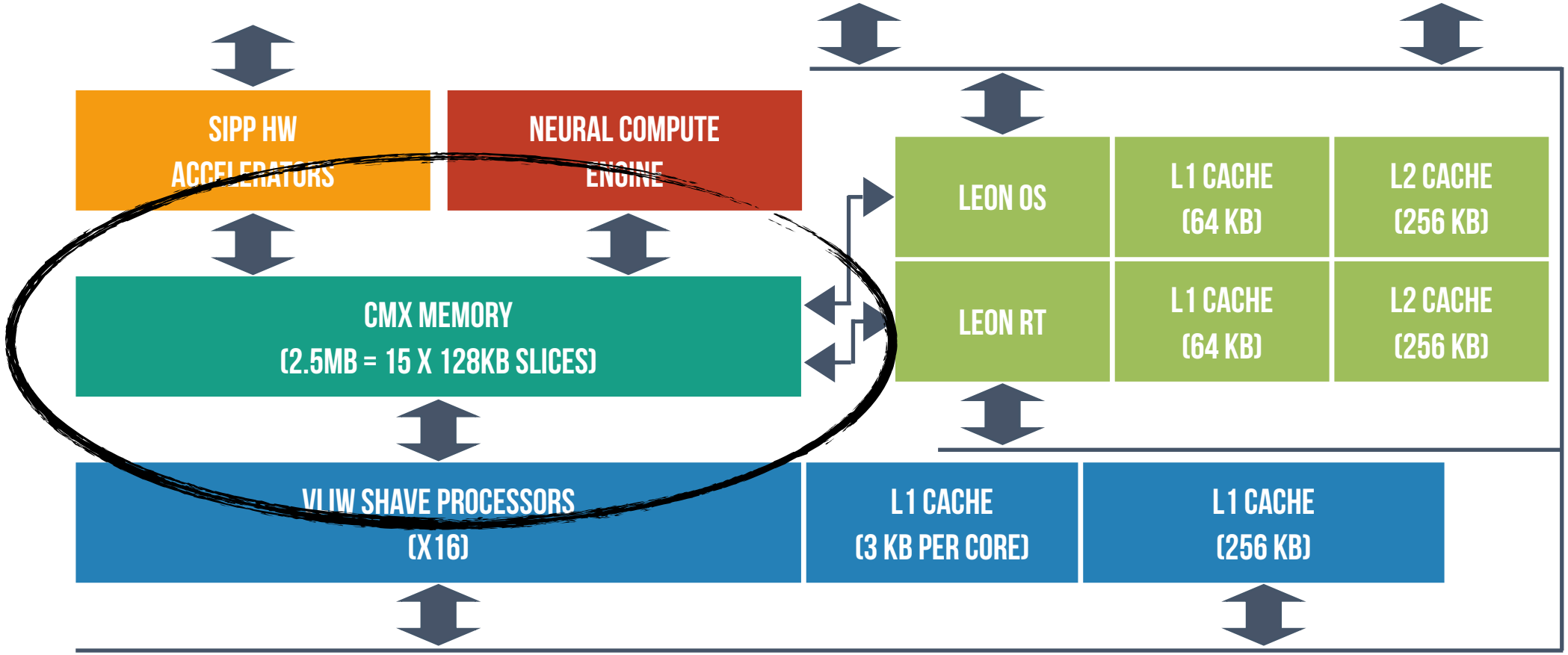
3. **Model Optimization**
   Difficult to perform model pruning.

4. **Matrix Format**
   Sparse matrix is not supported.

5. **Pruning Type**
   Structured pruning is only available.

INTEL MOVIDIUS MYRIAD X VPU HARDWARE

Petrongonas et al. (2021), ParalOS: A Scheduling & Memory Management Framework for Heterogeneous VPUs

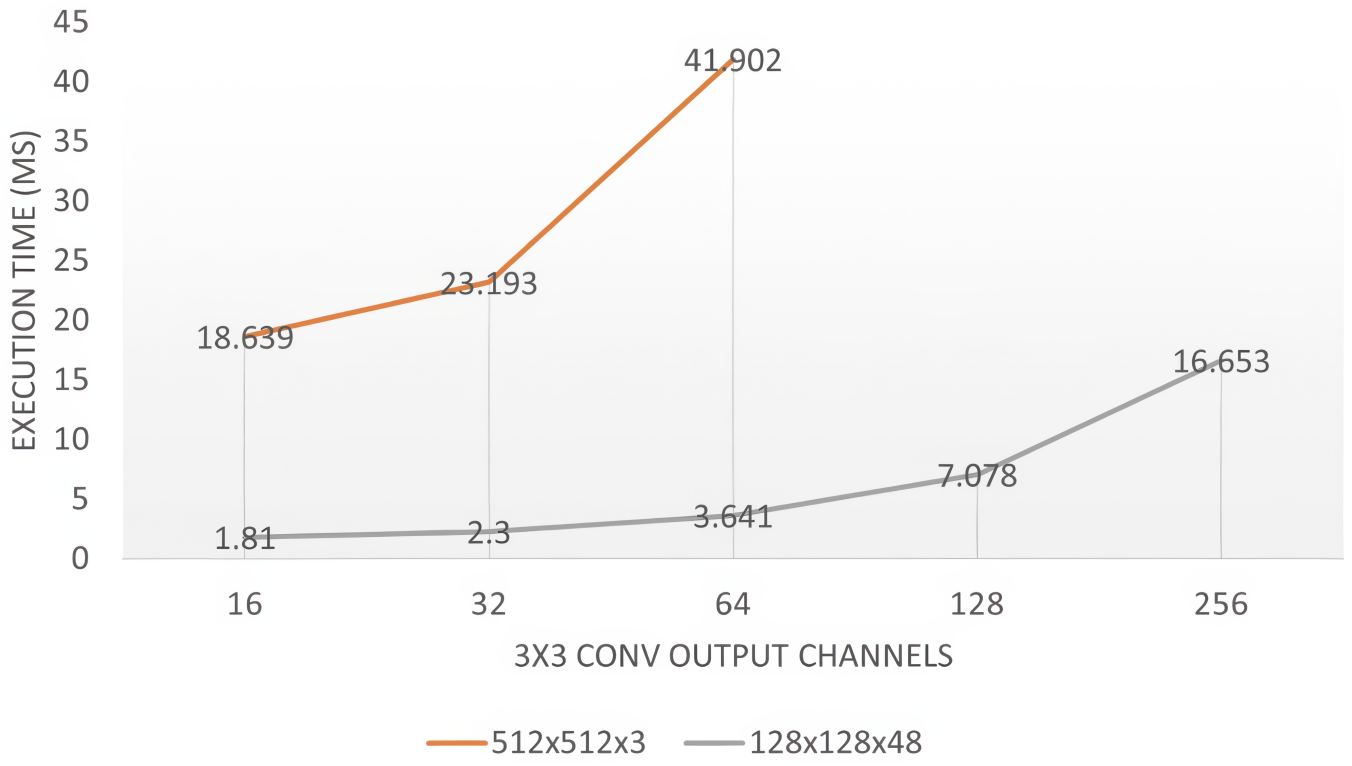


MODEL DESIGNING

Understanding Hardware

> Problem: Perform computation at higher spatial resolutions.

WHAT WE FOUND FROM EXPERIMENTS



SOLUTION

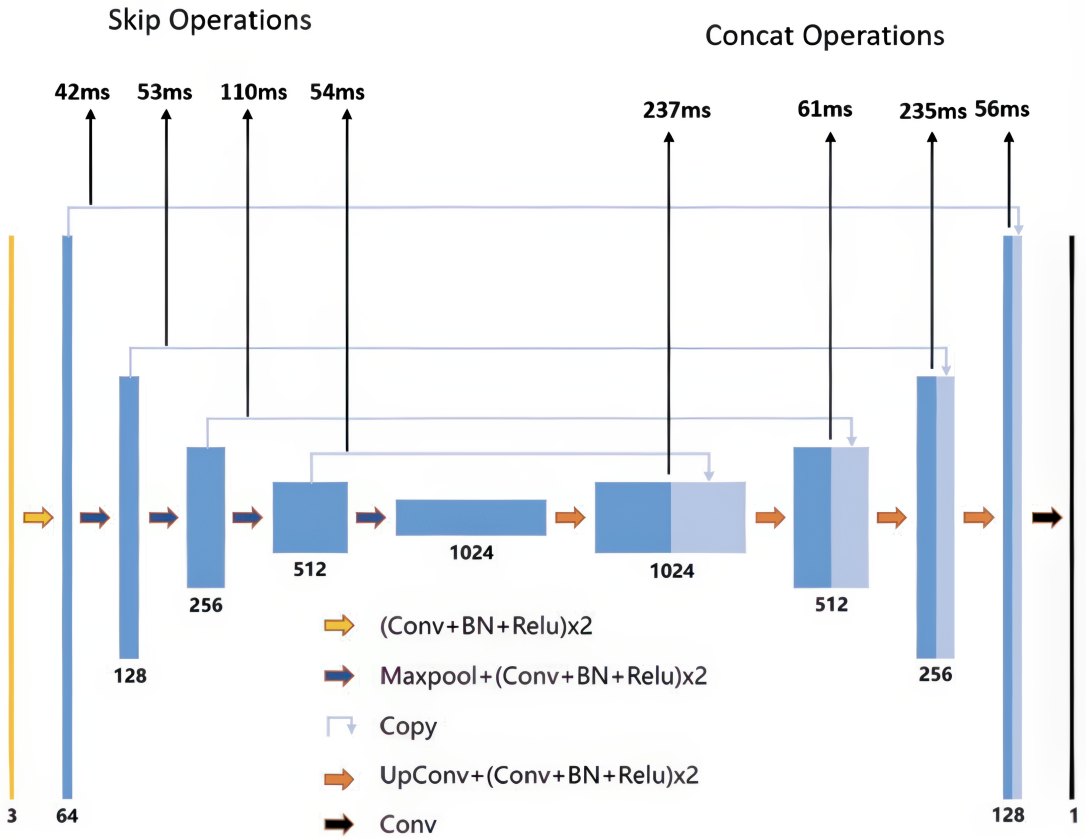
Use techniques such as pixel shuffling and large kernel convolution.

MODEL DESIGNING

Understanding Hardware

"Problem: Skip connection operations are costly if the feature size is large."

WHAT WE FOUND FROM EXPERIMENTS



SOLUTION

When a skip connection is used, reduce feature size.

MODEL DESIGNING

Understanding Hardware

Architecture

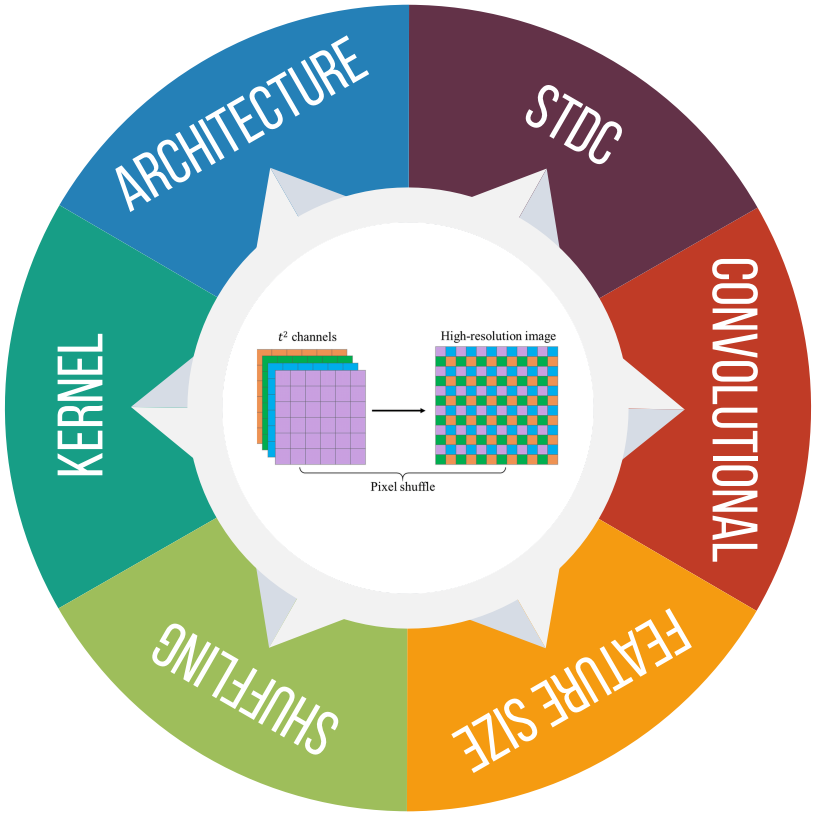
Making sure all the layers are executed using Neural Compute Engine (NCE)

Kernel

Use large kernel convolutions with large stride at input.

Pixel Shuffling

Use pixel shuffling at the output.



STDC (Convolution Block)

Use a modified Short-Term Dense Concatenate module (STDC).

Convolutional

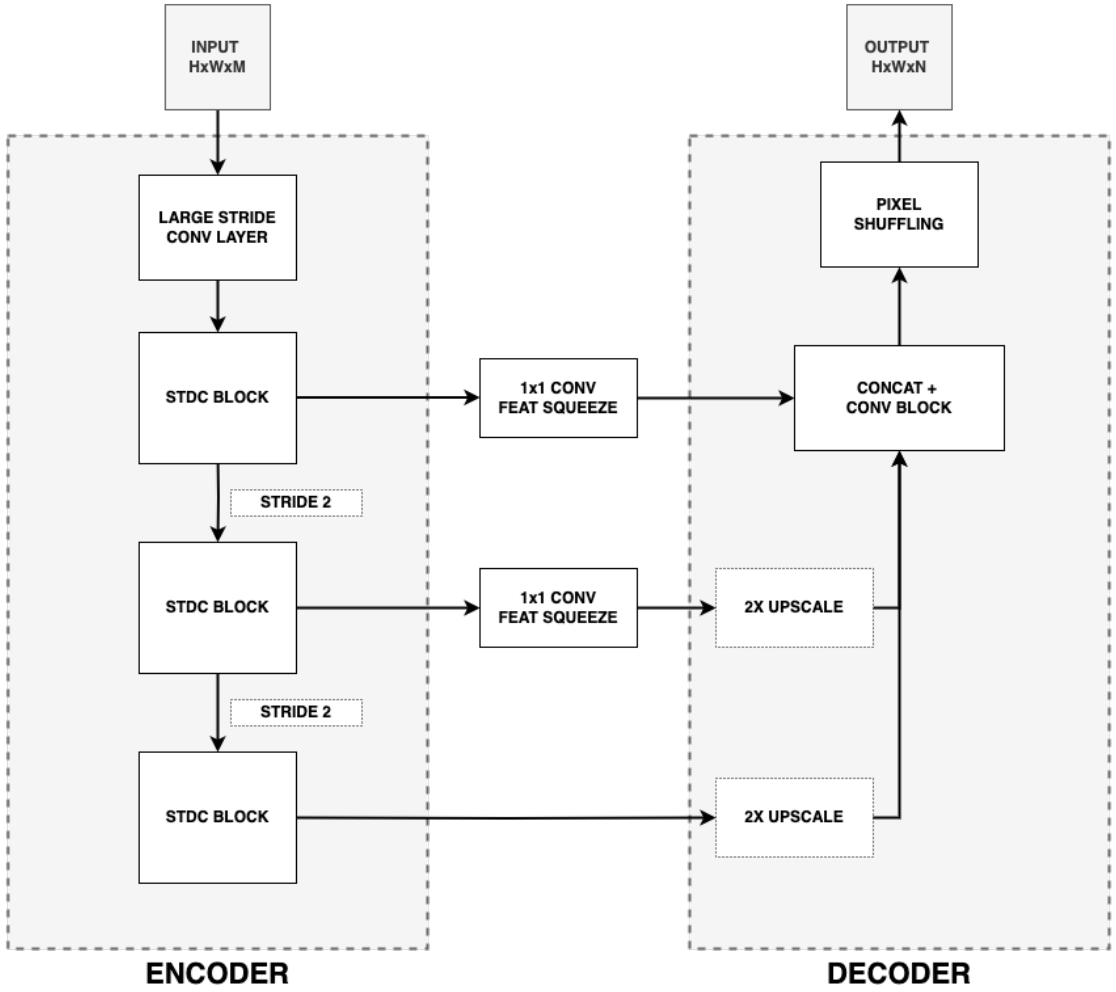
Use 1x1 CONV operations to reduce feature size for longer skip connections.

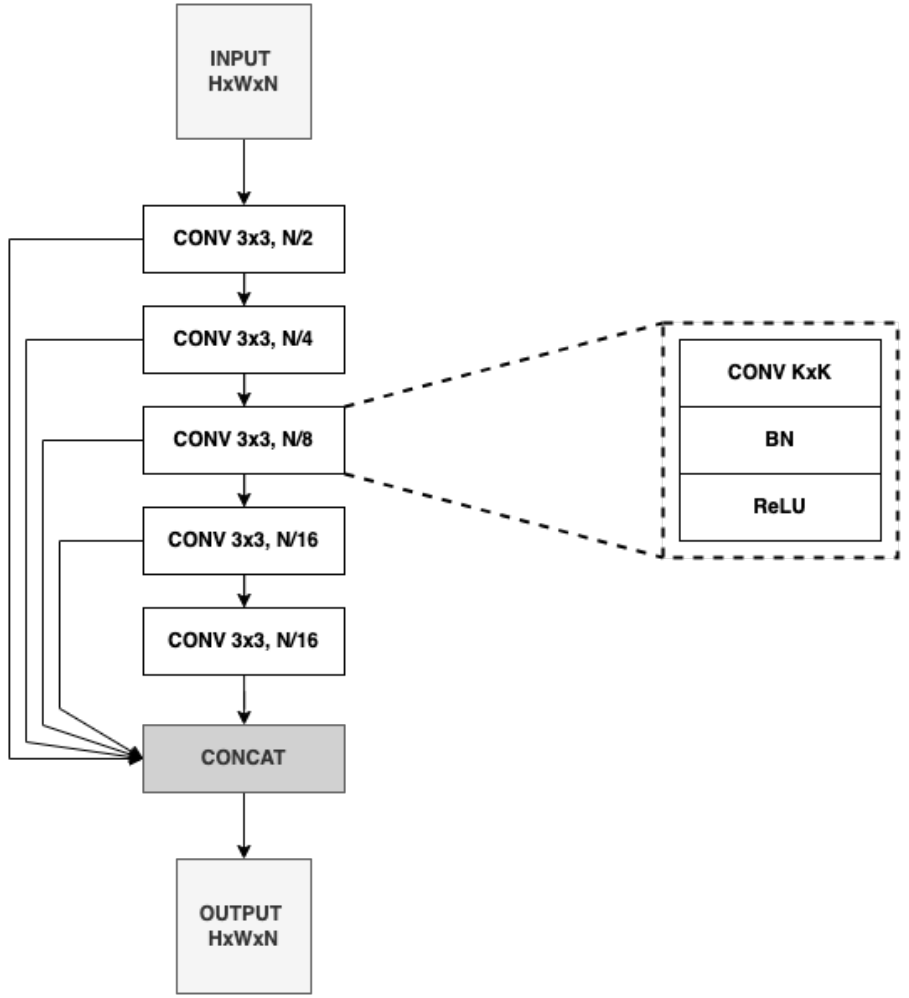
Feature Size

Use small feature size convolution layers to reduce copy-retrieve operations cost.

PROPOSED MODEL DESIGN

Mingyuan et al. (2021), Rethinking Bisenet for Real-Time Semantic Segmentation



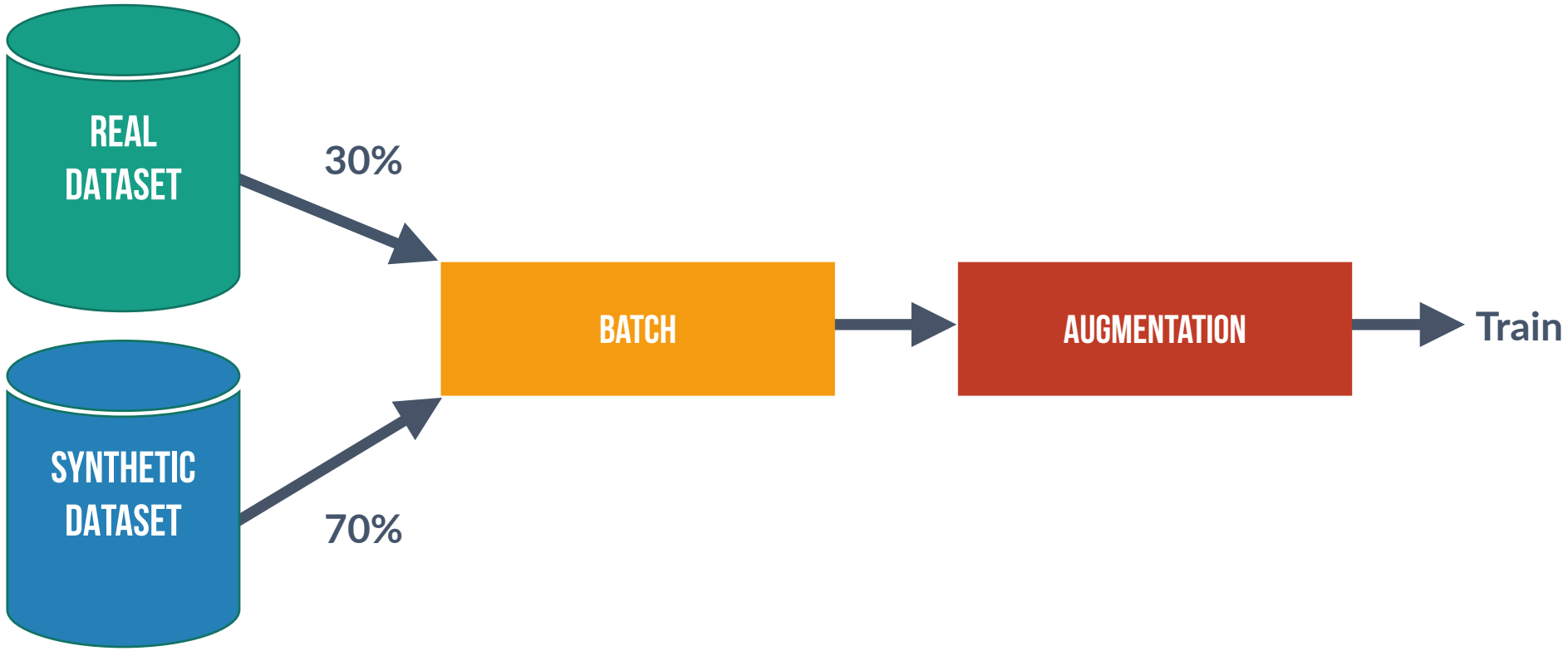



REAL IMAGES DATASET
Body Segmentation Datasets


SYNTHETIC IMAGES DATASET
Body Segmentation Datasets

DATASETS PIPELINE

Sample distribution for each bath



DATASETS AND TRAINING

Pipeline

DINO Pre-Training

It refers to a method of pre-training deep learning models using self-supervised learning techniques.

Training batch

For each batch, we feed 30% real data and 70% synthetic data.

Adam Optimizer

We used the 1e-4 learning rate in the Adam optimizer to update the model parameters.

Number of Batch

We set the number of Batch to 10K to train our segmentation model.


BODY SEGMENTATION RESULTS
The model runs at 18ms on hardware

BODY SEGMENTATION RESULTS

Comparison

Jabra PanaCast 20
On-Device Background Segmentation

Unified Communication
Platform 01

Unified Communication
Platform 02

QUESTIONS & ANSWERS

BREAK (30 MINUTES)

THANK YOU!