Yuhua Cheng

digital-nomad-cheng | Yuhua Cheng | Nomad's Wonderland |

Result driven High Performance Computer System master student with four years experience working on computation intensive computer vision projects on edge devices. Excellent problem solving skills with open mindset working in a team.

Education

2022 - present	Master - High Performance Computer Systems at **Chalmers University of Technology**
2014 - 2018	Bachelor - Electrical Engineering at **Beijing Normal University**
06/2023 - 07/2023	PUMPS+AI summer school on programming and tuning massively parallel computers at **Barcelona Super Computing Center**
2015/07 - 2015/09	Summer School at **University of California Berkeley**

Work Experience

Machine Learning Enginner at Verisure June 2024 - now

- Data engineering with AWS, Computer Vision model training, pruning, quantization, adapt model format for NPU.

Master Thesis at SmartEye Jan 2024 - Jun 2024

- Developing abandon object detection and notifying algorithm combining computer vision and deep learning for vehicle cabin on embedded devices.

Summer Intern at Zenseact Jun 2023 - Aug 2023

- Using CUDA to parallelize trajectory generation algorithm, achieved 20x to 600x speedup depends on scenario.
- Developing algorithm for generating optimal autonomous lane change trajectory while avoiding dynamic obstacles.
- Developing algorithm for generating extreme maneuver for emergency collision avoidance within maximum steering capability while keeping the computation cost low.

Senior Computer Vision Engineer at Meican Group Aug 2021 - Apr 2022

- Define the action pattern categories and data collection pipeline in our unmanned shelves.
- Research into and implement real time video action recognition solutions such as TSM using training framework PyTorch and deploy with efficient inference framework MNN.

Computer Vision Engineer at iHandy Group July 2018 - July 2021

- Implement various image processing algorithms such as face smooth, light correction. Familiar with basic image processing, feature extraction and pattern matching algorithms on embedded devices.
- Research into real time face detection/tracking, landmark localization, portrait matting, style transfer algorithms, and implement them on our Android/iOS mobile devices. Familiar with neural network design, training, optimization(knowledge distillation, pruning, quantization), and deployment.
- Research into image classification such as NSFW recognition and face attribute classification algorithms and deploy them using TensorFlow Serving. Familiar with data collection and annotation pipelines.
- Responsible for improving the computation efficiency of our algorithm solutions such as choosing various inference engine such as TensorFlow Lite/CoreML/ncnn, optimizing neural network efficiency with software and hardware.

Computer Vision Intern at RedTea AI Jan 2018 - July 2018

- Responsible for designing fast road defects detection solutions using Caffe and deploying it on edge devices such as Raspberry Pi. Our solution has been tested on thousands miles of high way road of China and achieved recall rate of more than 99%.

Projects

Project Course on AI Compilers using TVM Link

Study the internals of machine learning compiler framework Apache TVM, including frontend, backend, BYOC etc. Benchmark performance of different codegen mechanisms. Using BYOC to support a highly optimized neural network inference library ncnn. Achieved 30% speedup in AlexNet than ARM Compute Lib on Raspberry Pi 4B.

Face Detection on edge device using TVM Link

Implement the RetinaFace detection solution using C++ with the help of TVM. Benchmark the TVM auto schedule performance under small object detection network. The runtime library performance from TVM auto shcedule is on pair with MNN. Show how to parse input and output of TVM for object detection tasks.

CUDA Kernel Auto Schedule using TVM Link

Use TVM's auto schedule for generating efficient gemm(general matrix multiplication) kernel for Nvidia GPU, it outforms the human hand tuned kernel 2x on GFLOPs.

Fast Iris Landmarks Localization Link

Implement a real time iris landmarks localization solution using PyTorch framework. it can be used on all kinds of application such as driver drowsiness detection.

Face Detection on Edge Device using MNN Link

Implement the RetinaFace detection solution using C++ with the help of MNN. The solution can run on top various backend such as CPU or GPU with a simple switch. All kinds of solutions can be built on top of this such as face detection/tracking, and face recognition.

Face Portrait Matting on iOS Link

Implement a fast portrait matting algorithms solution, the model is trained using PyTorch, the inference is done using MNN, and the deployment is on iOS.

Hair Segmentation Service Link

Implement some navie hair segmentation neural networks using various dataset and build a hair color calculation service utilizing TensorFlow Serving.

Camera ISP pipelines Link

Implement various image signal processing(ISP) pipelines and computational photography algorithms utilizing auto-schedule library Halide.

Publications

Wang, Yikun et al. (2018). "Text2Sketch: Learning Face Sketch from Facial Attribute Text". In: *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 669–673. doi: `10.1109/ICIP.2018.8451236`.

Skills

Programming Language	C/C++, Python, CUDA, OpenMP, MPI, , OpenGL/ES, go, Objective-C
Tools	docker, git, vim
Libraries	Apache TVM, TensorRT, OpenCV, PyTorch, TensorFlow, Halide, MNN, ncnn, TensorFlow Lite, CoreML, LLVM, MediaPipe, Detectron2, MMDetection, gem5
Specializations	GPU Programming, Machine Learning Compilers, Deep Learning, Computer Vision, Image Processing, Computer Graphics, Neural Network on Edge Devices, High Performance Computing, Computational Photography, Distributed Systems

Last updated: April 14, 2025