ML Accelerator for Autonomous Driving Algorithm

Ritarka Samanta, Connor Talley, Minseung Jung, Gayoung Nam, Dinyar Islam, Owen Kew
Advisor: Dr. Callie Hao

Georgia Institute of Technology

School of Electrical and Computer Engineering

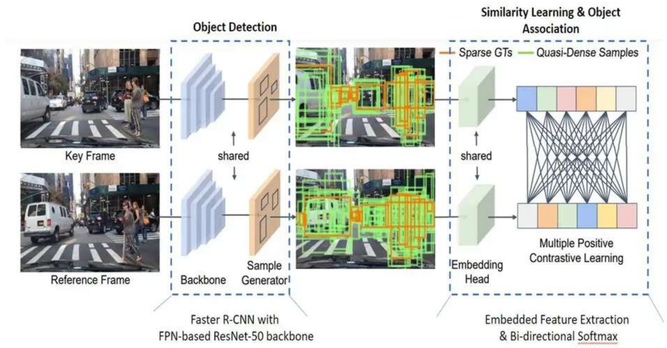
Motivation

Object Tracking algorithm is quite slow

HLS and FPGAs can be used to design complex hardware

Use FPGAs to accelerate object detection algorithm

Overview

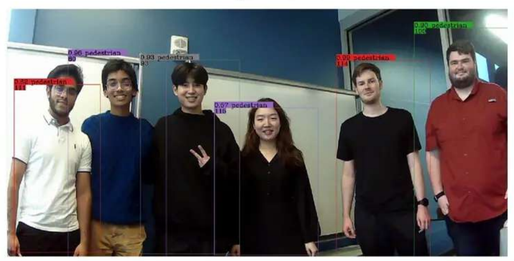


Source: Quasi-Dense Similarity Learning for Multiple Object Tracking, CVPR 2021

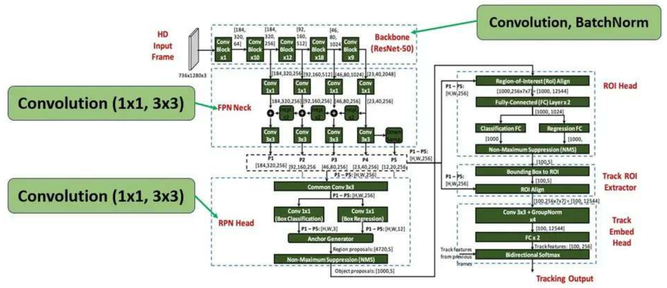
Detect multiple objects in a HD frame
Track objects across frames
Accurately, Efficiently, In real-time!

Dataset	FPS
MOT17	25-30
MOT20	25
BDD100K	5
Waymo	10
TAO	1

QDTrack

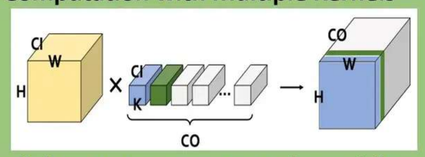


Block diagram for FPGA Implementation



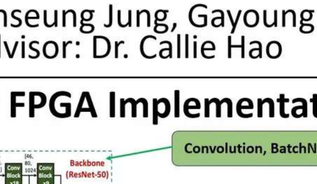
How to optimize the HLS code

Method 1:
CNN Computation with Multiple Kernels



Parallelism 1: Across output channel (kernel)
Speedup: CO
Parallelism 2: Across input channel
Speedup: CO * CI
Parallelism 3: Across width dimension
Speedup: CO*CI *W

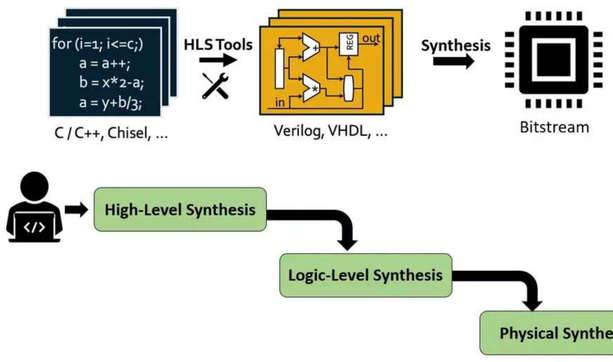
Method 2:
Tile Parallelization



Parallelism 1 : Intra-tile parallelization involves optimizing convolution computation

Parallelism 2 : Inter-tile parallelization is achieved through multiple kernel instantiation

How to Synthesize the HLS code

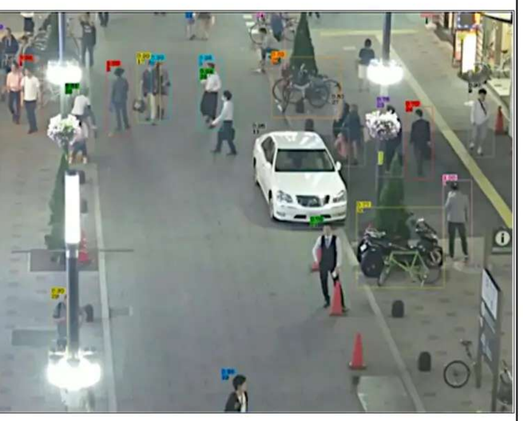


ResNet50 layer's latency

run on Xilinx ZCU

Layer Name	Latency (ns)	BRAM	DSP	FF	LUT
0.0.1	3.117e+12	5488 (300%)	26 (1%)	3273 (~0%)	7012 (2%)
0.0.2	2.894e+10	5209 (300%)	13 (~0%)	4523 (~0%)	9823 (3%)
1.0.0	5.529e+09	214130 (11739%)	15 (~0%)	10142 (1%)	10754 (3%)
1.0.1	1.448e+09	427570 (23441%)	17 (~0%)	10275 (1%)	9885 (3%)
1.0.2	2.452e+09	322349 (17672%)	17 (~0%)	10242 (1%)	10235 (3%)
...	...	...	...	...	...

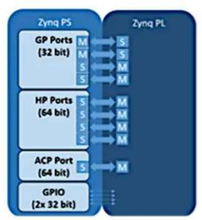
Live Inference for QDTrack



Future Work

1. Reduce BRAM Utilization

2. Implement inter-FPGA Communication



3. Synthesize entire algorithm into hardware