

Autonomous Intelligent Drones

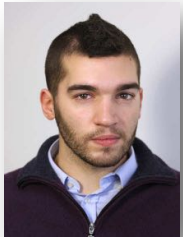
Christos-Savvas Bouganis

Christos-savvas.bouganis@imperial.ac.uk

Intelligent Digital Systems Lab

Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl



Stylianos I. Venieris
Machine Learning



Alexandros Kouris
Machine Learning,
Robotics



Konstantinos Boikos
Computer Vision,
SLAM



Manolis Vasileiadis
Computer Vision



Mudhar Bin Rabieah
Machine Learning



Nur Ahmadi
Brain-Machine Interface



Christos-Savvas Bouganis
iDSL Lab Director
Imperial College London

Intelligent Digital Systems Lab (iDSL)

Home About us Research Group members Publications Work with us Contact

Welcome to the Intelligent Digital Systems Lab at Imperial College

TOP LINKS

- Our research
- Dr. Christos Bouganis
- Join our lab
- CNN-to-FPGA Benchmark Suite
- fpgaConvNet

The iDSL lab is part of the Electrical and Electronic Engineering Department of Imperial College London.

Cascade^{CNN}: Pushing the performance limits of quantisation

Alexandros Kouris
Dept. of Electrical and Electronic Eng.
Imperial College London
a.kouris@imperial.ac.uk

Stylianos I. Venieris
Dept. of Electrical and Electronic Eng.
Imperial College London
stylianos.venieris@imperial.ac.uk

Christos-Savvas Bouganis
Dept. of Electrical and Electronic Eng.
Imperial College London
christos-savvas.bouganis@imperial.ac.uk

ABSTRACT

This work presents CascadeCNN, an automated workflow that pushes the quantisation limits of any given CNN model, to perform high-throughput inference by exploiting the computation time-accuracy trade-off. Without the need for retraining, a two-stage architecture tailored for any given FPGA device is generated, consisting of a low- and a high-precision unit. A confidence evaluation unit is employed between them to identify misclassified cases at run-time and forward them to the high-precision unit to increase computation. Experiments demonstrate that CascadeCNN achieves a performance boost of up to 33% for YOLO-10 and 68% for AlexNet over the baseline design for the same resource budget and accuracy.

1. INTRODUCTION

While Convolutional Neural Networks are becoming the state-of-

Figure 1: High-level CascadeCNN workflow

CNN device pair to select quantisation scheme, configure the confidence evaluation mechanism and generate the cascaded low- and high-precision processing units.

Research

In the Intelligent Digital Systems Lab, we perform research towards high-performance (embedded) digital systems spanning several topic areas, including machine learning, computer vision, and robotics.

MORE DETAILS

@CBouganis

Tweets by @CBouganis

Christos @CBouganis
We are recruiting for an exciting post on Machine Learning and FPGAs. Please see details here: tinyurl.com/yachfwmd

Research Assistant/Associat...
The Intelligent Digital Systems ...
imperial.ac.uk

Aug 4, 2018

Christos @CBouganis
Liked on 08/08/2018

Vision: Autonomous Intelligent Drones

Goals:

- Perceive the environment
- Understand the environment
- Interact with the environment

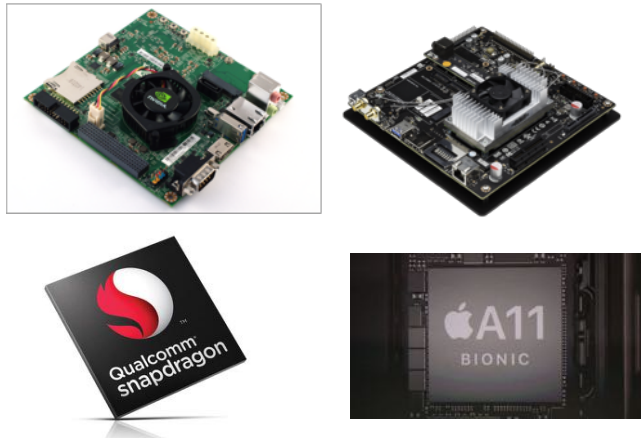
Challenges:

- Low latency
- Low power
- Adaptation



Conventional and Unconventional Embedded Platforms for Compute

GPUs – Tegra K1, X1 and X2
DSPs – Qualcomm Hexagon,
Apple Neural Engine, ...



- ✓ High throughput
- ✗ Low latency
- ✗ Low power
- ✓ Tools



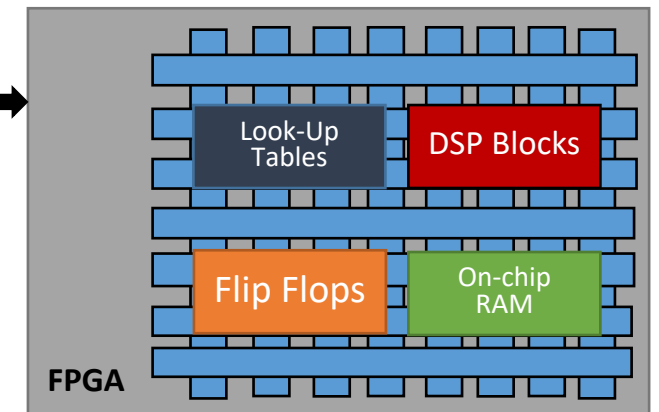
customisation

Ambarella
Myriad X

FPGAs

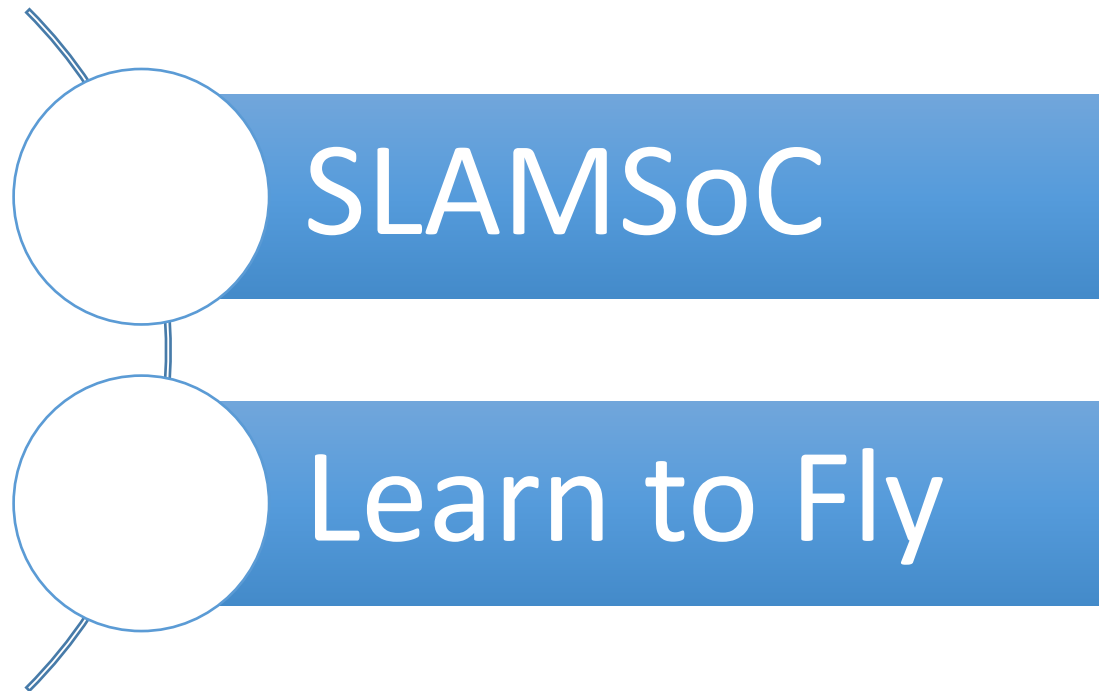
- Custom datapath
- Custom memory subsystem
- Programmable interconnections
- Reconfigurability

External Memory (DRAM)

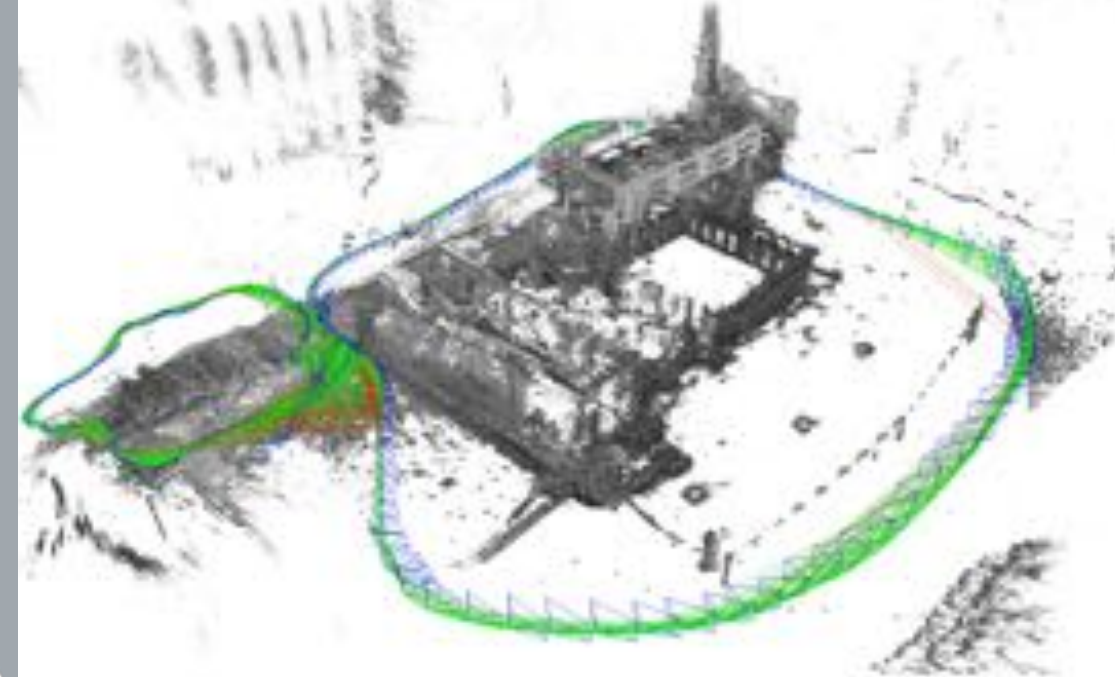


- ✓ High throughput
- ✓ Low latency
- ✓ Low power
- ✗ Tools

Challenge: Huge design space
Our Approach: Automated toolflows

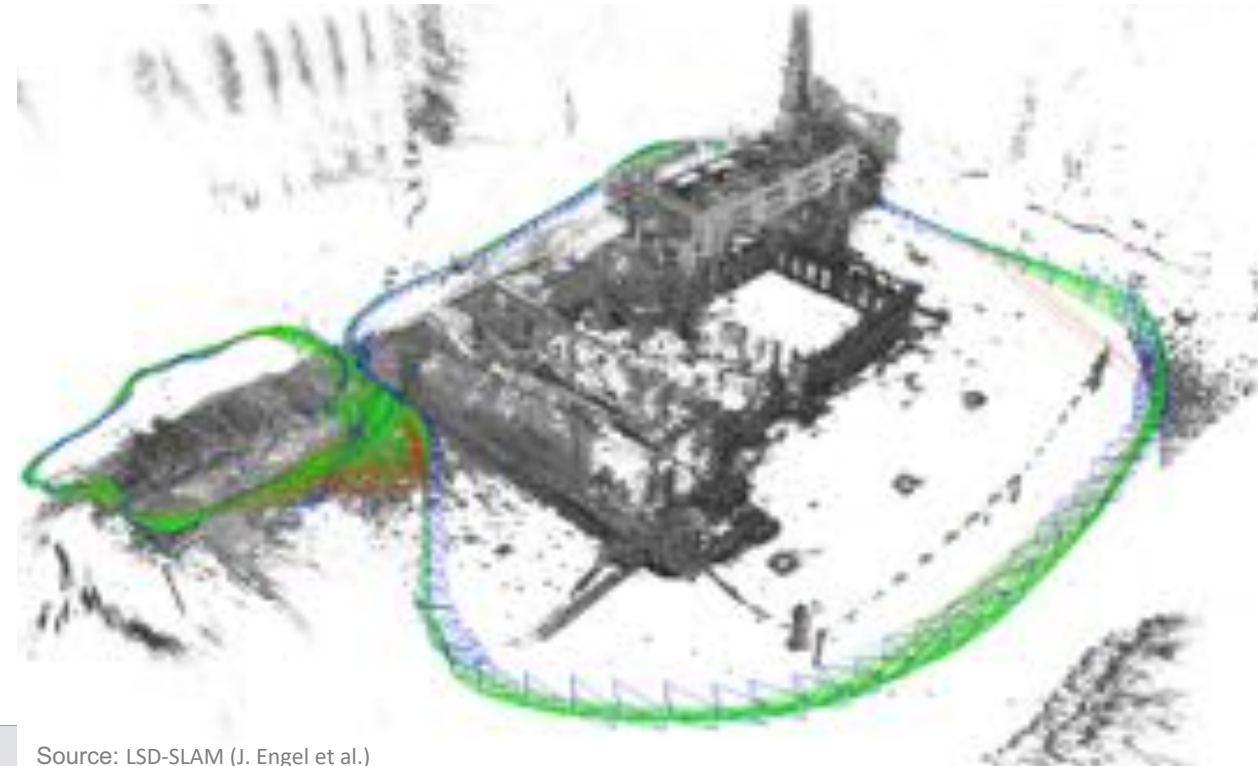
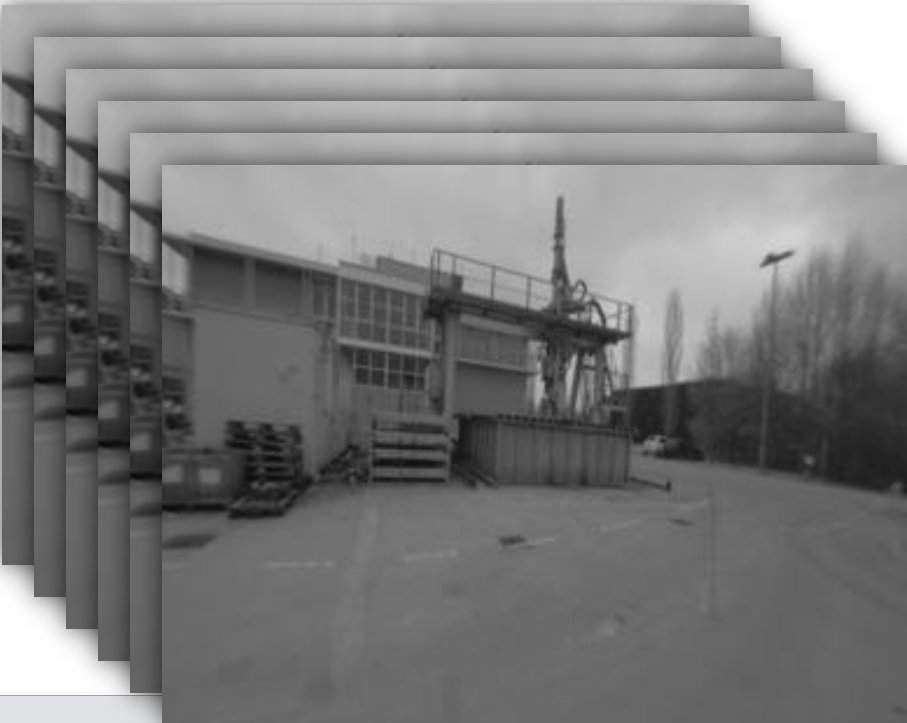


Topic #1: SLAMSoC



Use a series of observations to simultaneously perform Localisation and Mapping

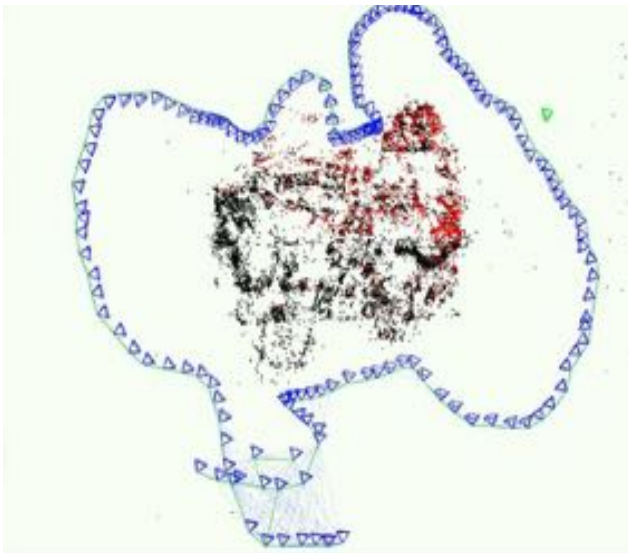
- Tracking (Localisation): Online pose estimation of the sensor and robot.
- Mapping: Fuse observations into a coherent model of the environment



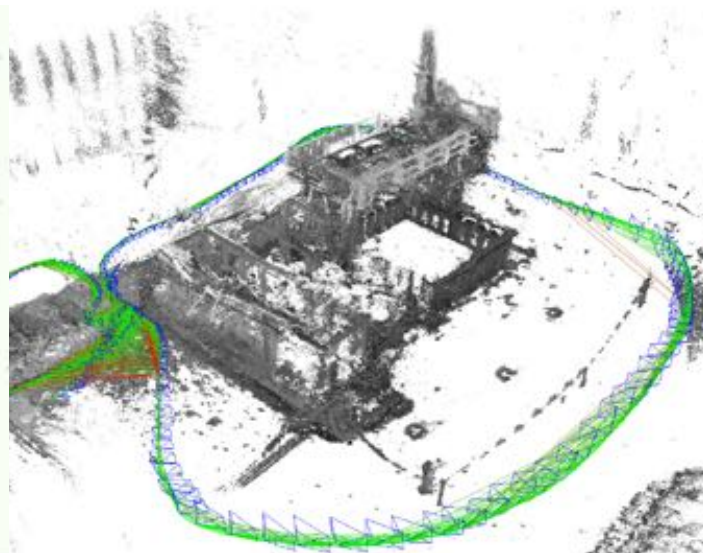
Source: LSD-SLAM (J. Engel et al.)

Challenges in Embedded SLAM

- Emerging algorithms have high complexity and bandwidth requirements
- Field still in a state of constant change
- **Tracking** robustly needs high framerate and low latency



Sparse
Mobile CPU

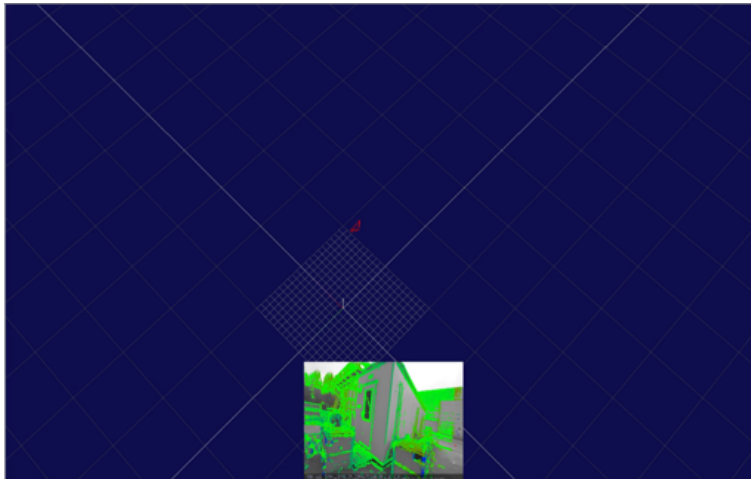


Semi-Dense
High-end Desktop

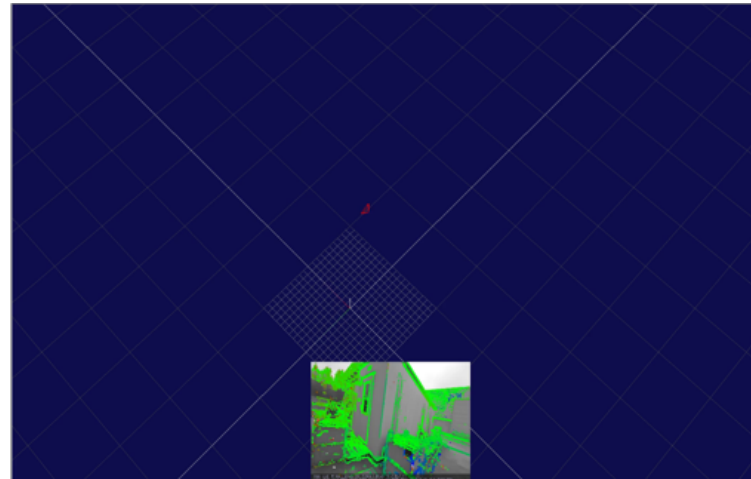


Dense SLAM
High Performance GPU Acceleration

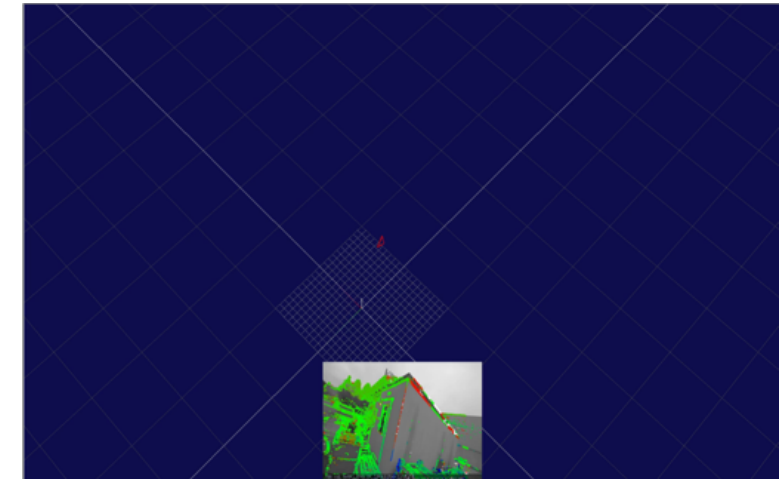
Importance of High Performance Tracking



- Camera rate (30fps)
- Intel i7-4770



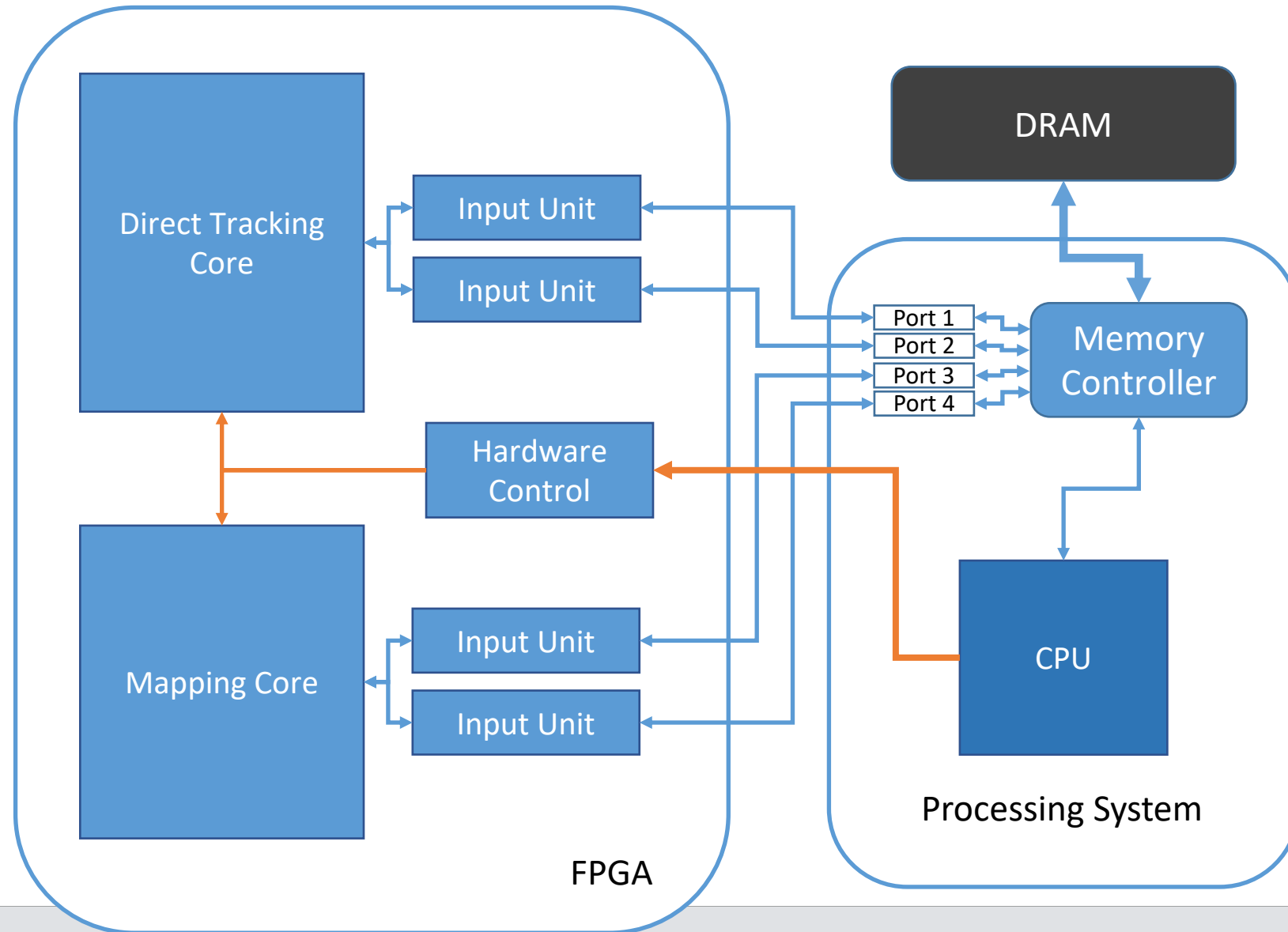
- Drop frames (15 fps processed)
- Position Drift, Error accum.



- Processing <8 frames/s
- Lost tracking

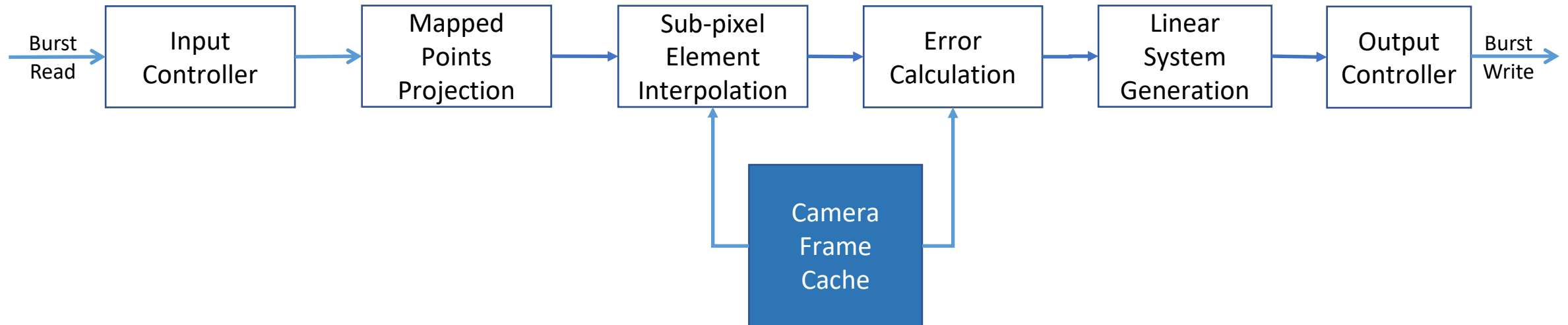
Proposed system

- Common memory space and Direct Memory Access
- Hardware high-level control from CPU
- Both operate simultaneously
- Buffered high memory bandwidth



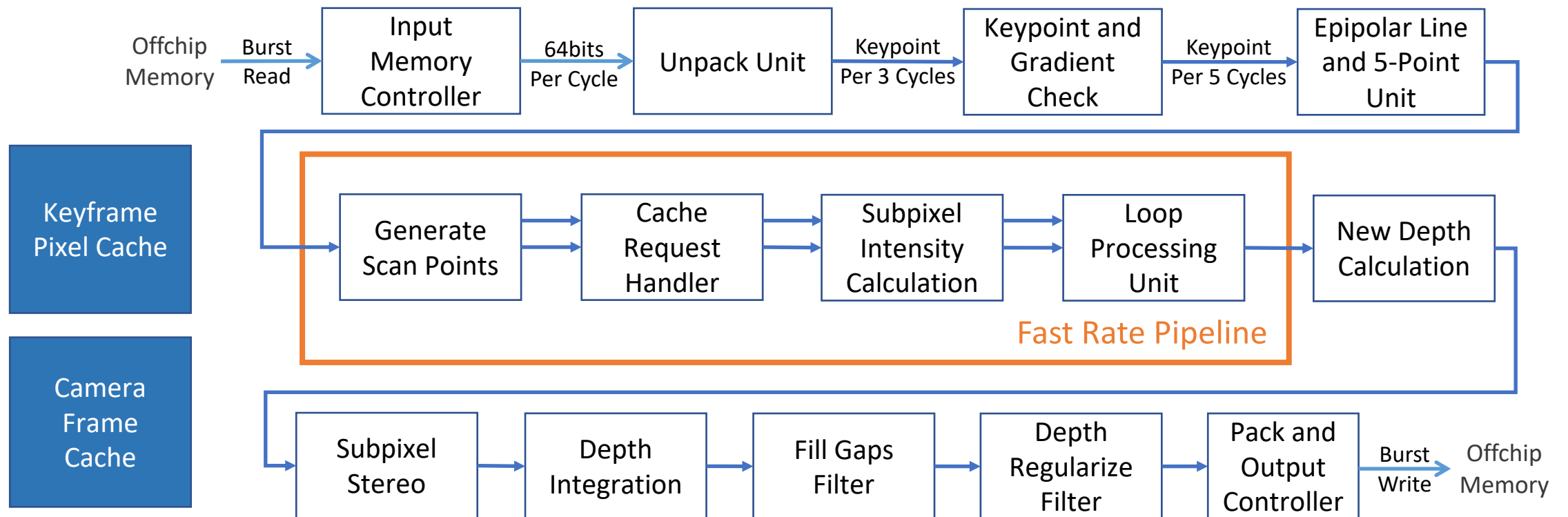
Direct Tracking Core

- Streaming Dataflow – Designed with High Level Synthesis
- Splitting computation into smaller blocks allows better optimisation
- Separating control flow from computation leads to a better design
- Redundant computation proved more efficient than going to memory

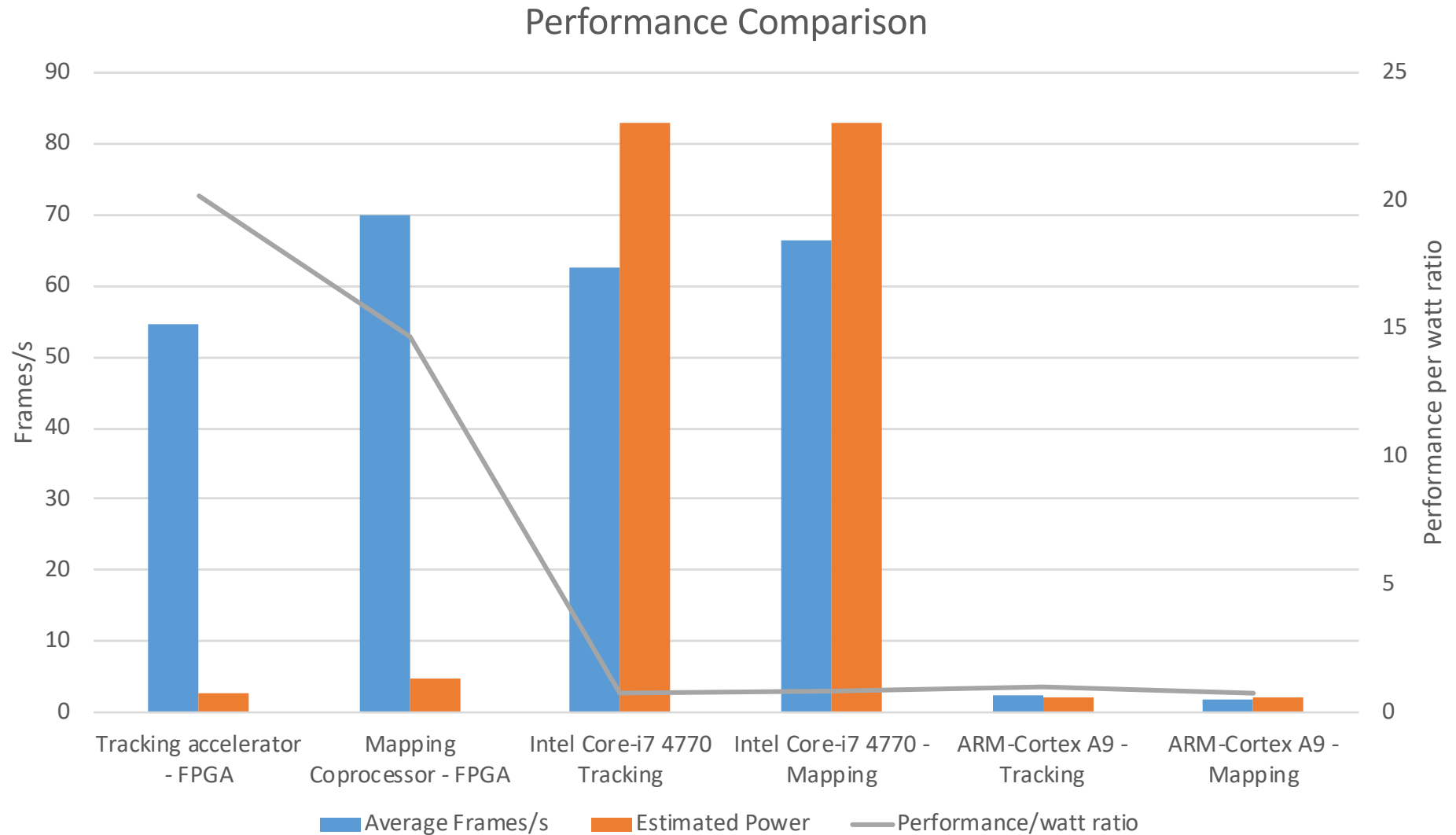


Mapping Coprocessor

- Variable rate pipelines
- Streaming dataflow processing, combined with local caches for random-access patterns



Comparison with other platforms



Topic #2: Learn to Fly



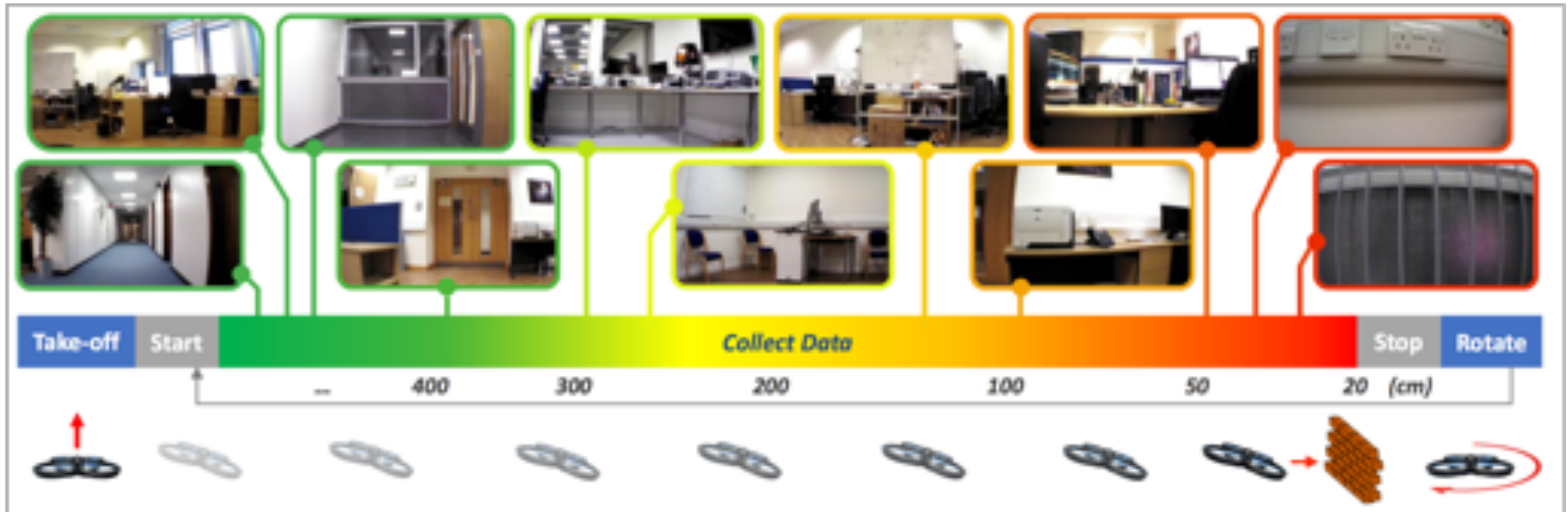
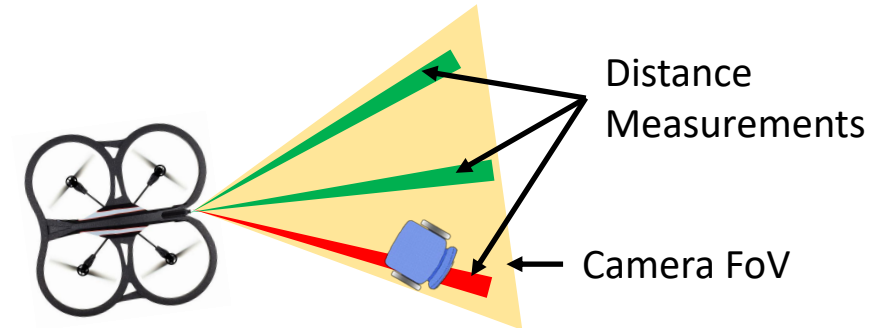
Self-Supervised approach for autonomous navigation:

- Exploits solely on-board camera's visual input
- Regression CNN to predict distance-to-collision
- Local path planner to modulate velocities



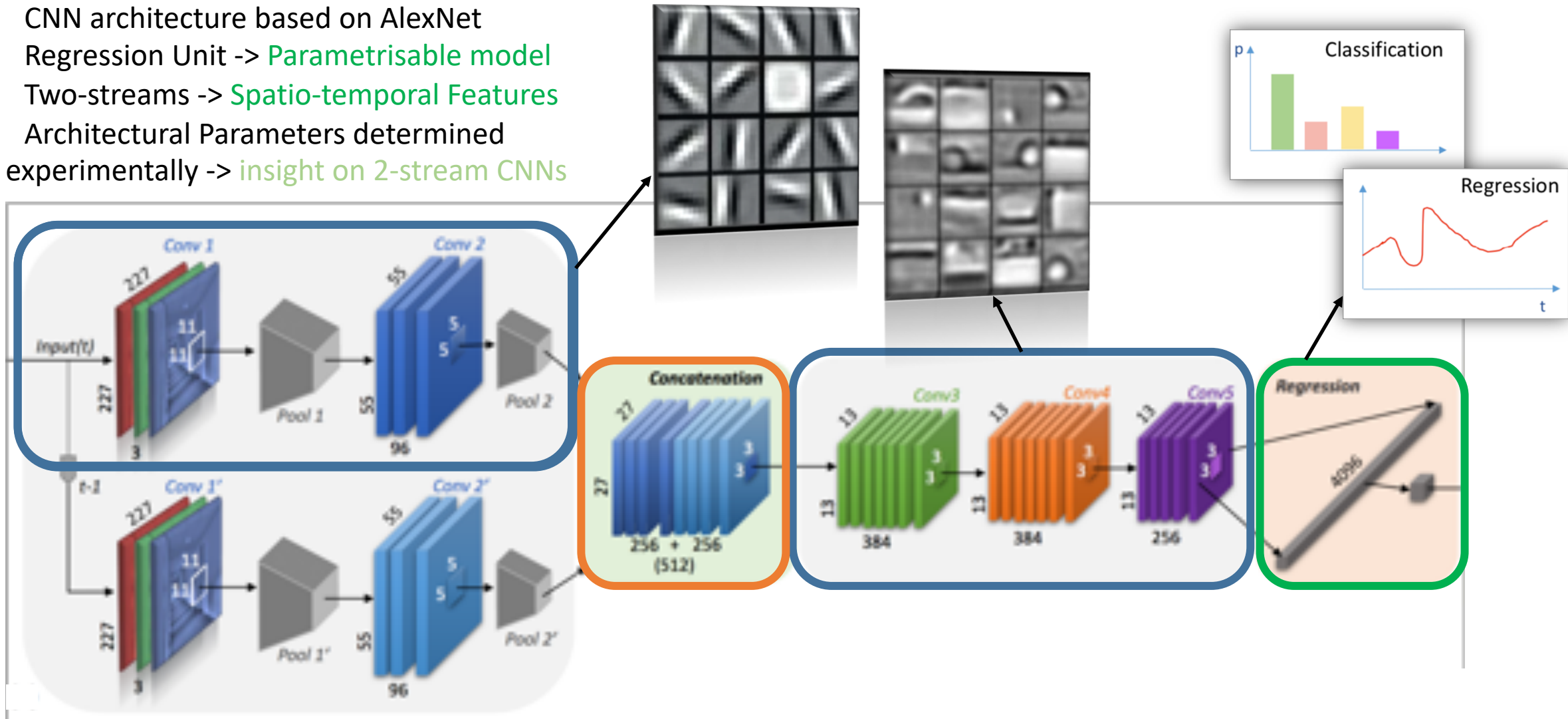
Self-Supervised Data Collection & Annotation

- Robust Deep Learning models require tons of data
- External Distance Sensors to automate the collection
- Indoor Flight Dataset:
 - Annotated with real-distance values
 - 300.000 samples in 2000 trajectories



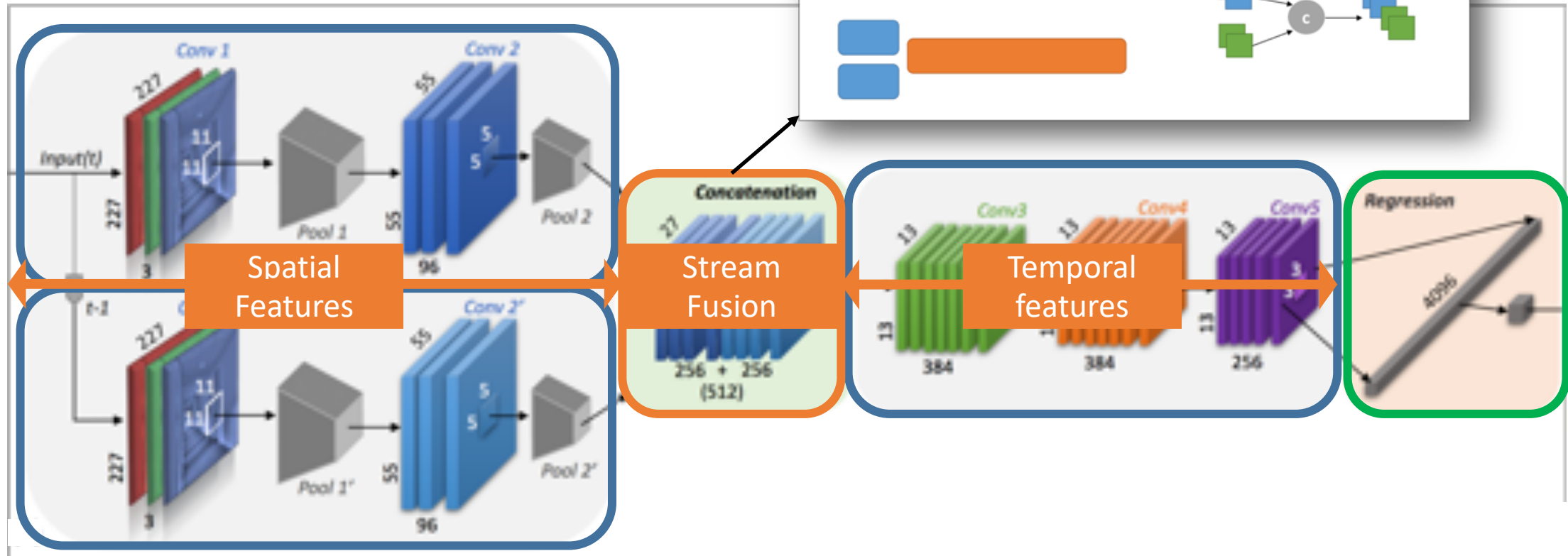
2-stream Regression CNN Architecture & Training

- CNN architecture based on AlexNet
- Regression Unit -> **Parametrisable model**
- Two-streams -> **Spatio-temporal Features**
- Architectural Parameters determined experimentally -> **insight on 2-stream CNNs**



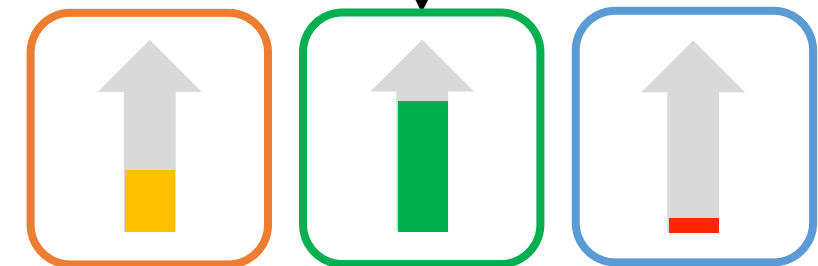
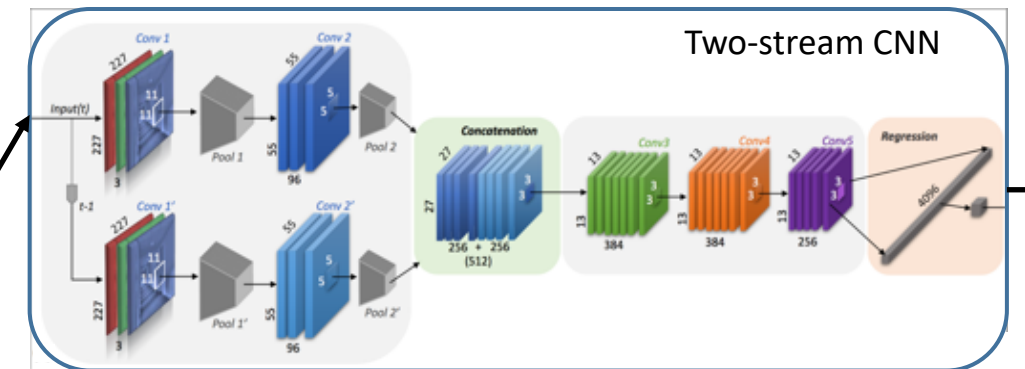
2-stream Regression CNN Architecture & Training

- CNN architecture based on AlexNet
- Regression Unit -> **Parametrisable model**
- Two-streams -> **Spatio-temporal Features**
- Architectural Parameters determined experimentally -> **insight on 2-stream CNNs**



2-stream Regression CNN Architecture & Training

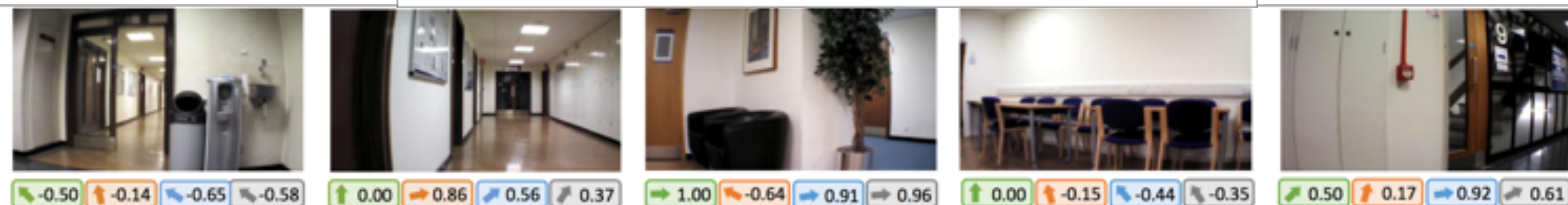
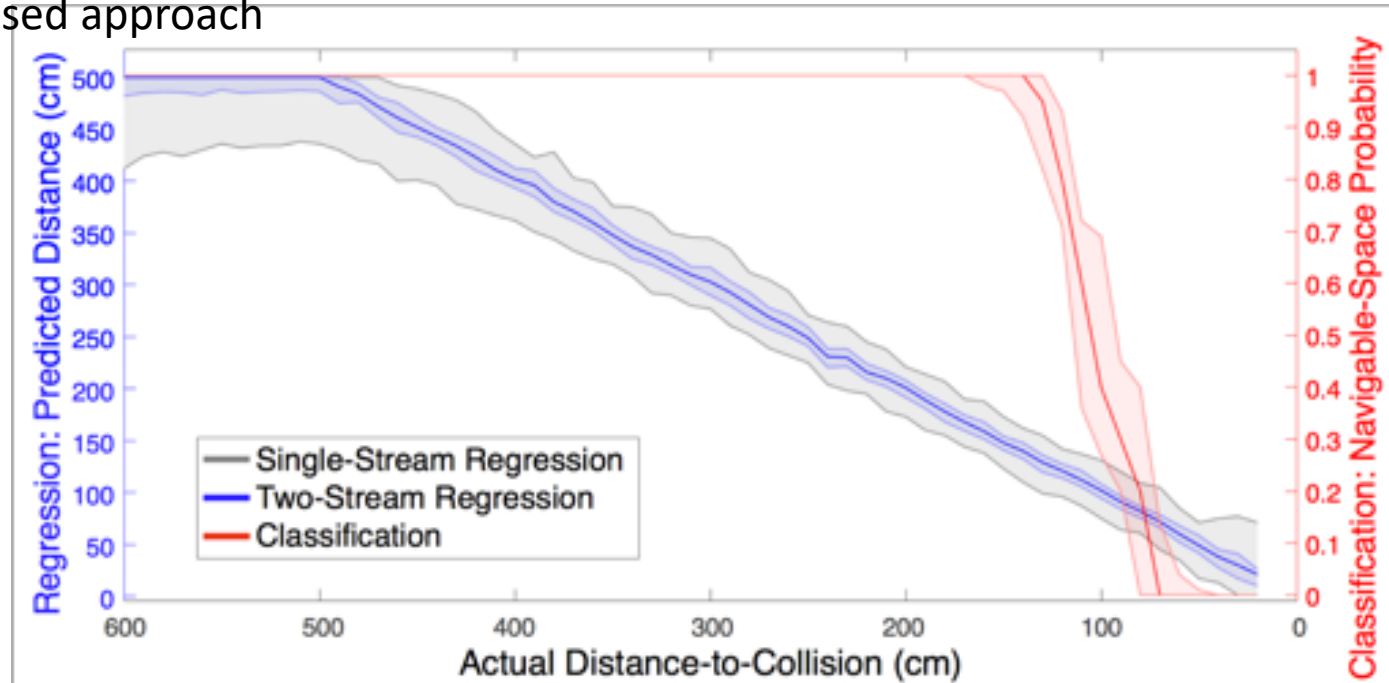
- The CNN predicts the distance-to-collision for three partly overlapping windows of the image



Distance-to-collision Predictions

Quantitative Evaluation: End-to-end UAV Navigation Task

Classification vs Regression-based approach



Normalised Yaw [-1,1]: Imitation Learning Classification Regression (This work) Human Pilot

