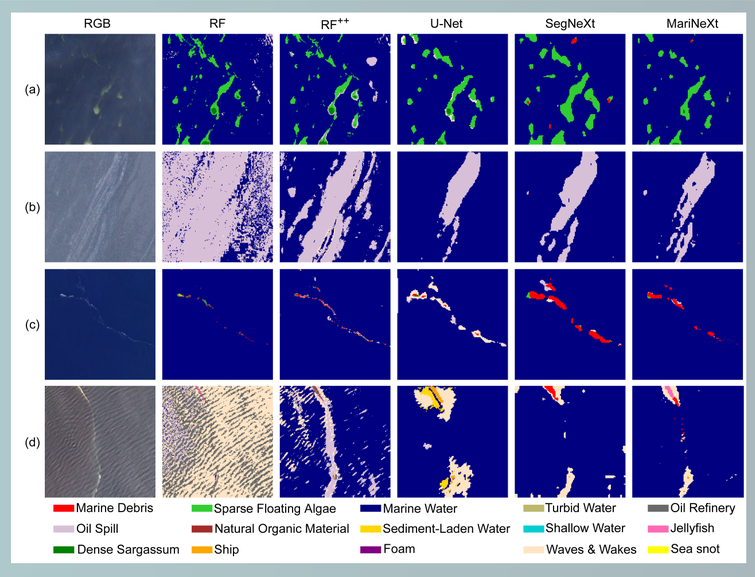
Edge SpAIce

Enabling Onboard Data Compression with Machine Learning on FPGAs

S. Tzelepis[1,2], N. Ghielmetti[1], N. M. Lemoine[3], M. Pierini[1], S. Summers[1], F. De Vielleville[3]

Oceanic Marine Litter Semantic Segmentation

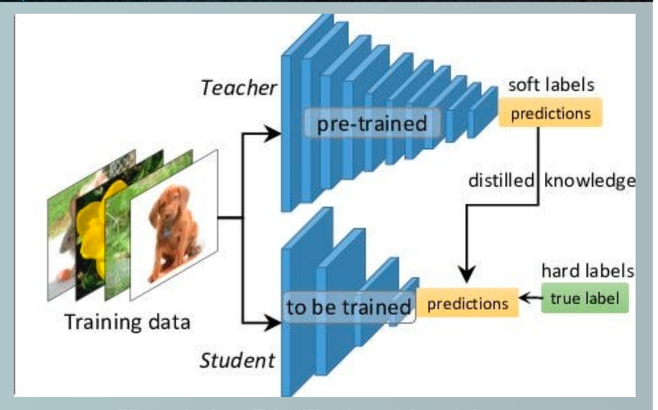
The project aims to enhance marine plastic litter detection by training **cutting-edge DNNs**. By integrating the **MADOS dataset** [1] and global datasets, NTUA team develops software for harmonising data, enabling more precise detection. The system distinguishes plastic litter from other sea surface features (waves, ships, etc.), offering more nuanced classifications models than basic binary classifiers. High-performance DNNs optimised for GPU clusters is trained to achieve superior detection accuracy, addressing scalability and generalisation challenges.



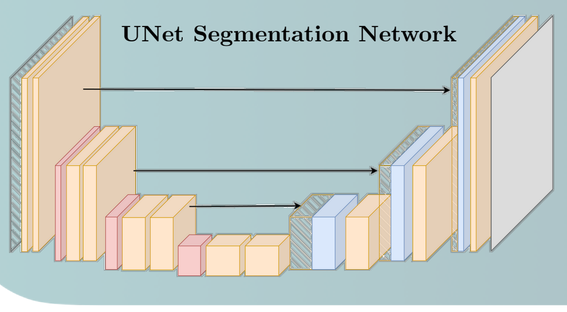
[1]: Detecting Marine pollutants and Sea Surface features with Deep learning in Sentinel-2 imagery

Quantisation-Aware Distillation

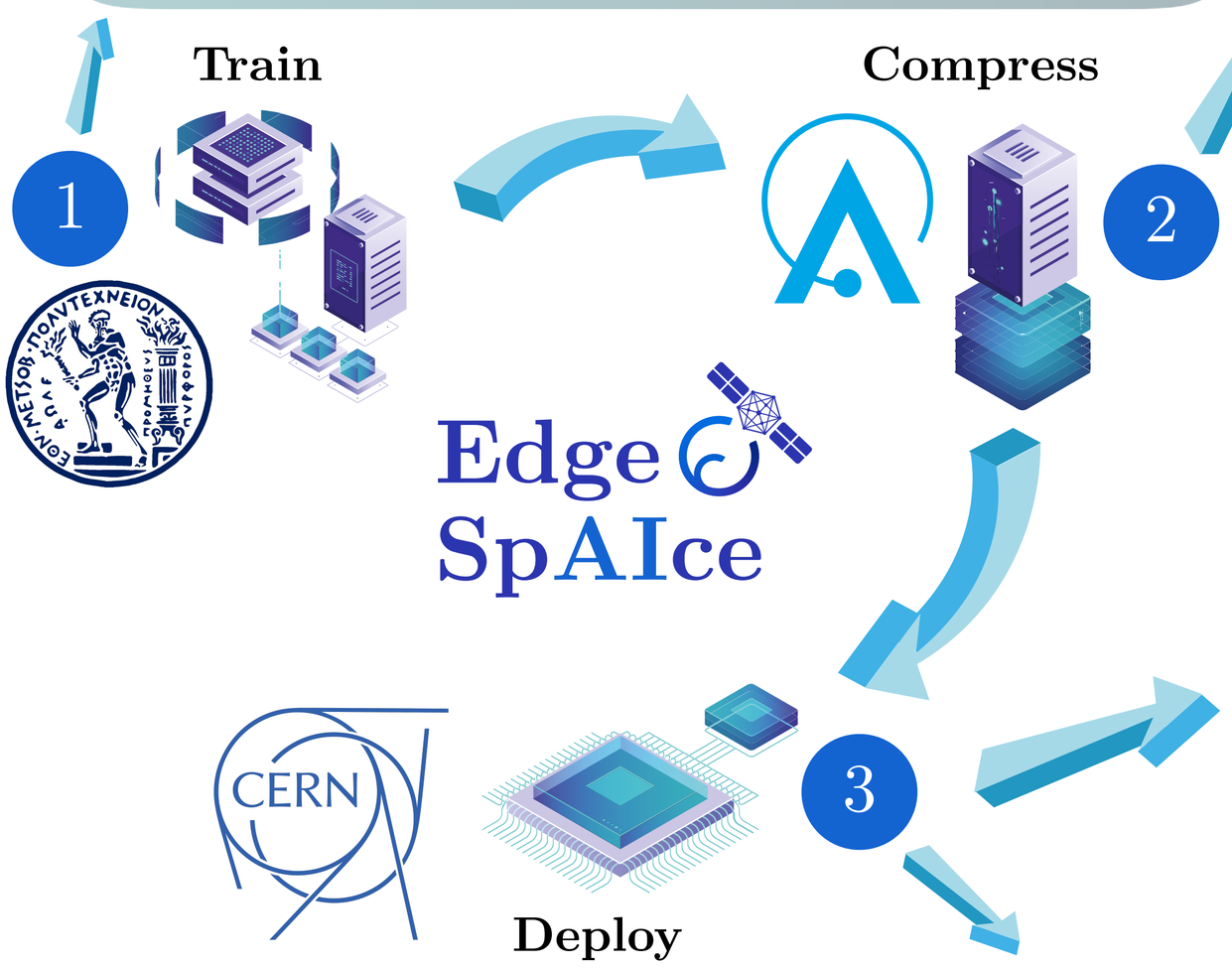
ODiToo, Agenium Space's proprietary software, is being enhanced within the Edge SpAIce project to **compress** large DNNs for deployment on Earth Observation satellites. The software enables DNN distillation, reducing network size to 1M or less parameters with **less than 6% accuracy loss** (3% from parameter reduction, 3% from quantisation).



Knowledge Distillation Algorithm

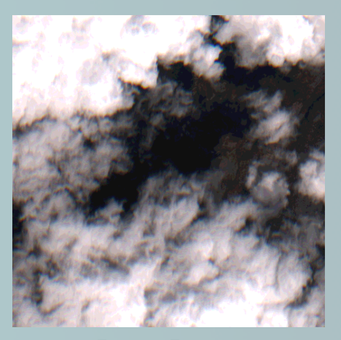


During training, **dynamic quantisation** will minimise accuracy loss. By using arbitrary bits per computation, ODiToo can support larger DNNs on the same FPGA size, maximising performance and scalability.

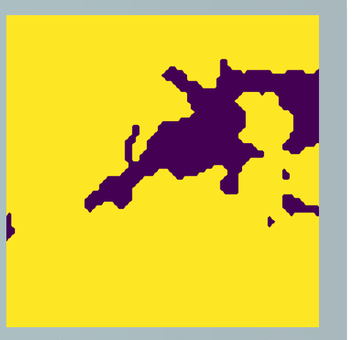


QONNX vs hls4ml

We demonstrate the pipeline of Knowledge Distillation, quantization and hls4ml deployment with a UNet model trained on the ALCD dataset for clouds segmenation [2]. The quantised and distilled model is exported in **QONNX format**, an open-source extension of ONNX developed by the Fast ML community to support arbitrary precision quantisation. In order to check the functional consistency between QONNX and the corresponding hls4ml generated model, the segmentation output produced by QONNX is compared with the one produced by hls4ml C-simulation.



Input Image



Segmentation mask

QONNX prediction

hls4ml prediction

[2]: Sentinel-2 reference cloud masks generated by an active learning method

FastML based FPGA implementation

FPGA Results

We measure the image processing framerate and power consumption. The **hls4ml** deployment of neural networks **outperforms Vitis AI**, achieving **8.8 times higher pixels per second per watt** on 50k parameters model. The performance gain is primarily due to hls4ml's **on-chip weight implementation**, reducing power consumption and memory bottlenecks compared to fetching from DDR memory.

Results [3] computed on ZCU102 SoC

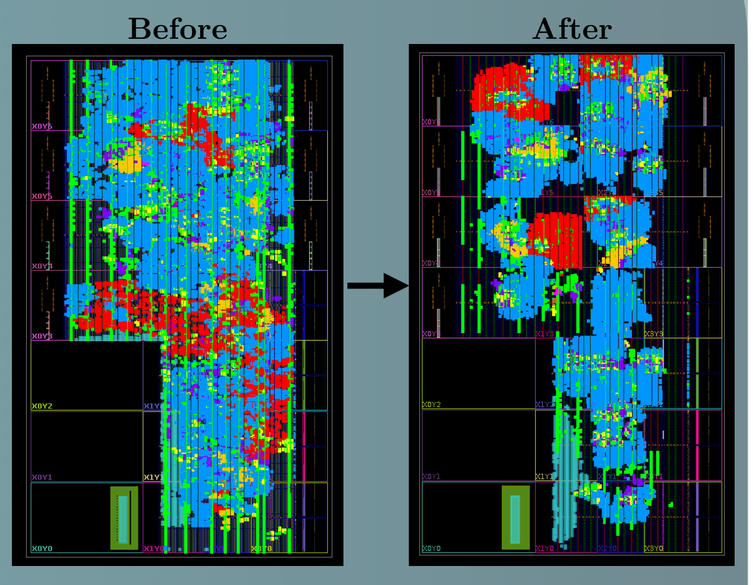
Platform	Framerate (FPS)	Power (W)	pixels/s/W($\cdot 10^6$)
`hls4ml`	834.3	3.9	14.2
Vitis AI	287.5	11.8	1.6
Ratio `hls4ml`/Vitis AI	2.9	0.3	8.8

Platform	LUT	LUTRAM	FF	BRAM	DSP
`hls4ml`	251396 (92%)	20323 (14%)	114324 (21%)	374 (41%)	246 (10%)
Vitis AI	163166 (60%)	20782 (14%)	303801 (12%)	769 (84%)	2144 (85%)

[3]: Edge SpAIce: Enabling Onboard Data Compression With Machine Learning On FPGAs

FIFO Depth Optimisation

hls4ml's dataflow architecture requires **FIFO buffers** between each NN layer to synchronise logic blocks. The FIFO depth depends on the latency and initiation intervals of the layers. An undersized FIFO slows down inference, while an oversized one wastes resources. To minimise resource usage while avoiding compute stalls, **FIFO depth optimisation [4] method reduces FIFO depths** to the minimum needed for each layer. After optimisation, most FIFO depths are **significantly smaller** compared to the initial values.



[4]: Real-time semantic segmentation on FPGAs for autonomous vehicles with hls4ml