IMPERIAL

Tiny End-to-End Collaborative Learning for Occlusion-Robust Object Detection

Presenter: Chieh-Tung (Danny) Cheng
17/09/2025

Agenda

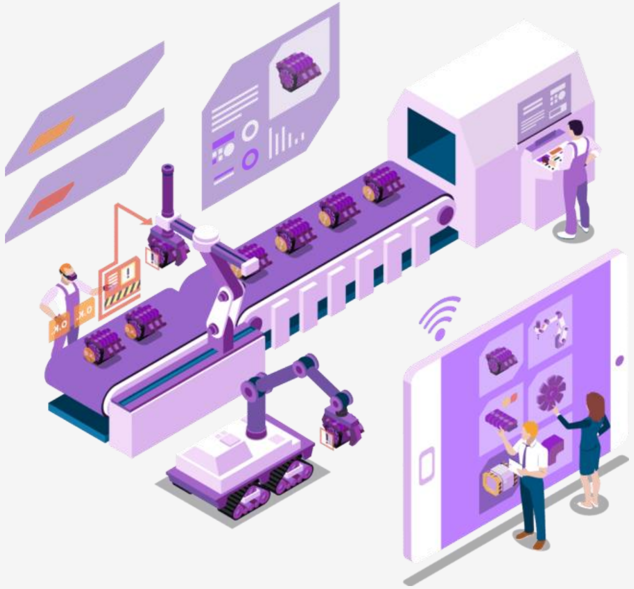
1 Background
2 Challenges
3 Related Work
4 Contribution
5 Experiments and Findings
6 Conclusion
7 Q & A

Background

The rapid growth of edge devices

- The economic potential of the Edge Devices and IoT is forecasted to reach $5.5 trillion to $12.6 trillion globally by 2030 [1]
- Computer vision (CV) plays a central role in Edge AI, driven by the widespread use of visual sensors at the edge.

Autonomous driving



Equipment Monitoring

Search and Rescue (SAR)

[1] Chui M, Collins M, and Patel M. The Internet of Things: Catching Up to an Accelerating Opportunity. Accessed: 2025-08-24. 2021. Available from: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/iot-value-set-to-accelerate-through-2030-where-and-how-to-capture-it

Challenges

Three main challenges in Edge AI (Ultra-low-end devices)

- **Resource constraints:** Lack GPUs, with strict limits on energy and memory
  - ➢ Motivates collaborative use of multiple edge devices.

- **Communication cost:** Overhead, synchronisation delays, and topology design can significantly impact performance and feasibility

- **Environment complexity:** Dynamic and visually challenging scenes make robust detection more difficult.

Related Work

Survey of TinyML, Fusion, and DFL, with Remaining Gap

Tiny Computer Vision Models

- MCUNet families

Model Compression Techniques

- Pruning, quantisation, knowledge distillation, neural architecture search

Collaborative Inference in CV

- Single model, Voting mechanism, Ensemble-based models

Robust CV Models under Occlusion

- Feature-level Fusion, Decision-level Fusion

Decentralised Federated Learning (DFL) in CV

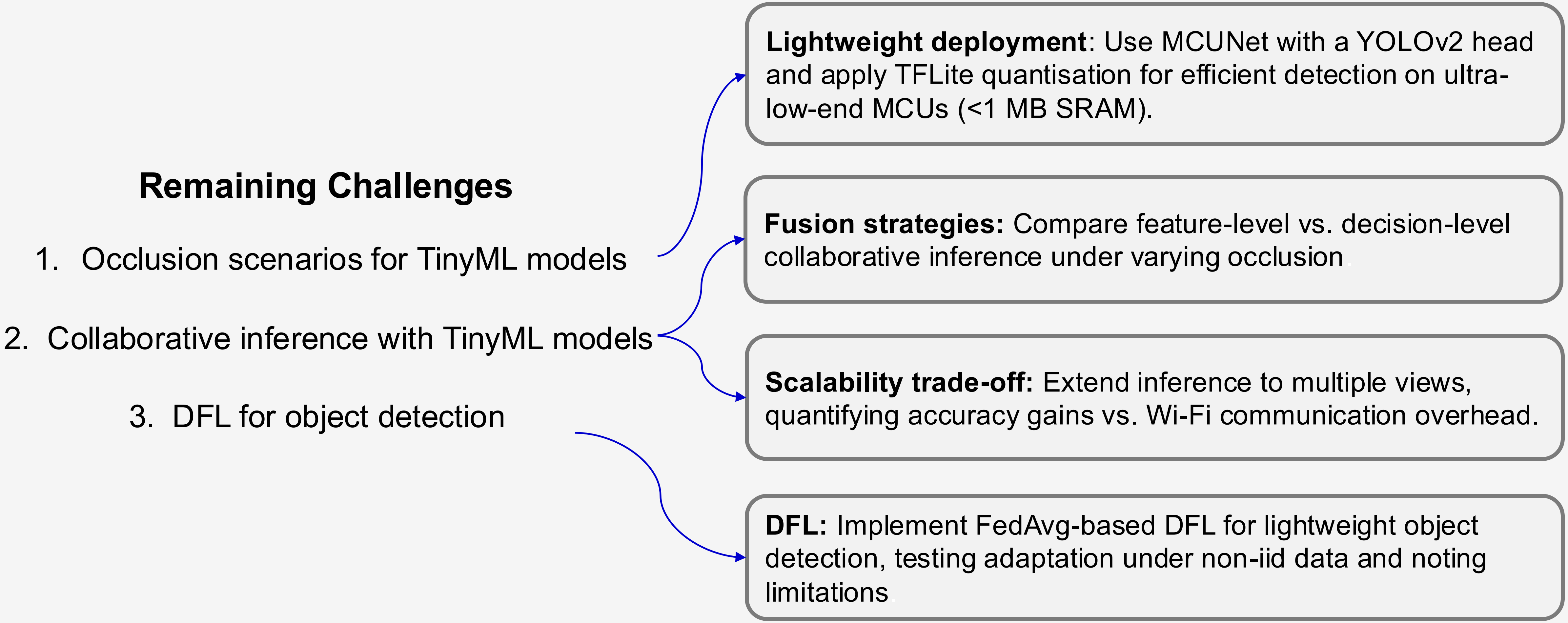
- network topology, communication protocol, and learning paradigm

Remaining Challenges

1. Occlusion scenarios for TinyML models
2. Collaborative inference with TinyML models
3. DFL for object detection

Contribution

How we address the remaining challenges



Remaining Challenges

1. Occlusion scenarios for TinyML models
2. Collaborative inference with TinyML models
3. DFL for object detection

Lightweight deployment: Use MCUNet with a YOLOv2 head and apply TFLite quantisation for efficient detection on ultra-low-end MCUs (<1 MB SRAM).

Fusion strategies: Compare feature-level vs. decision-level collaborative inference under varying occlusion.

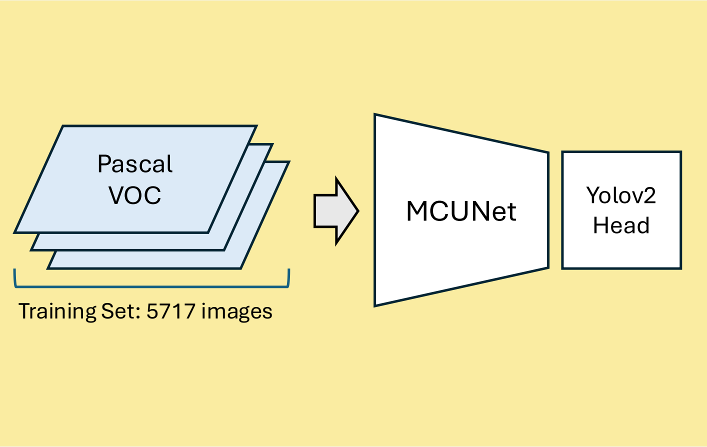
Scalability trade-off: Extend inference to multiple views, quantifying accuracy gains vs. Wi-Fi communication overhead.

DFL: Implement FedAvg-based DFL for lightweight object detection, testing adaptation under non-iid data and noting limitations.

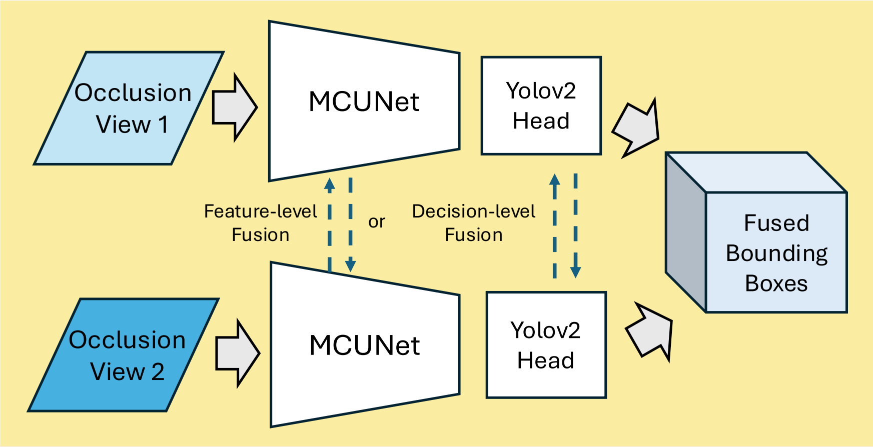
System Overview

Pre-training, Collaborative Inference and DFL

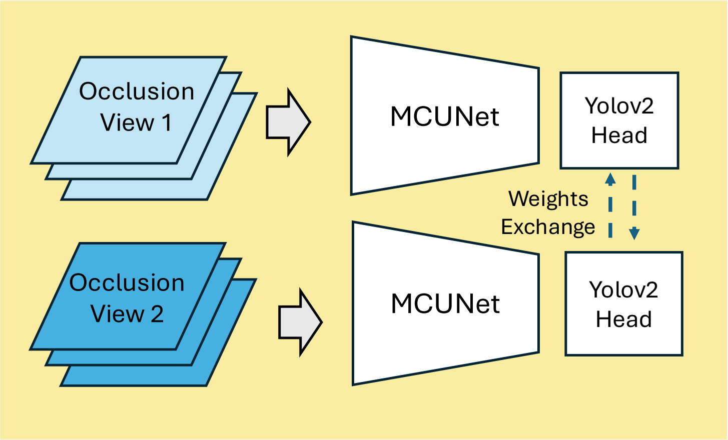
1. Pretraining



2. Collaborative Inference



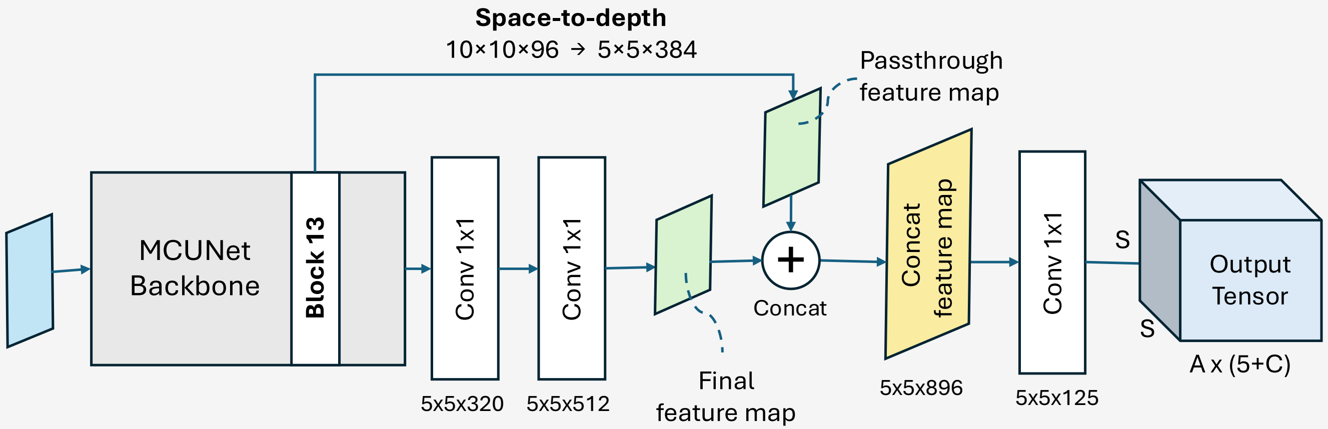
3. DFL



Experiment: Pretraining

Model / Dataset / Metric / Evaluation

Model



MCUNet backbone + YOLOv2 detection head

Dataset

- PASCAL VOC (object detection benchmark)

Metric:

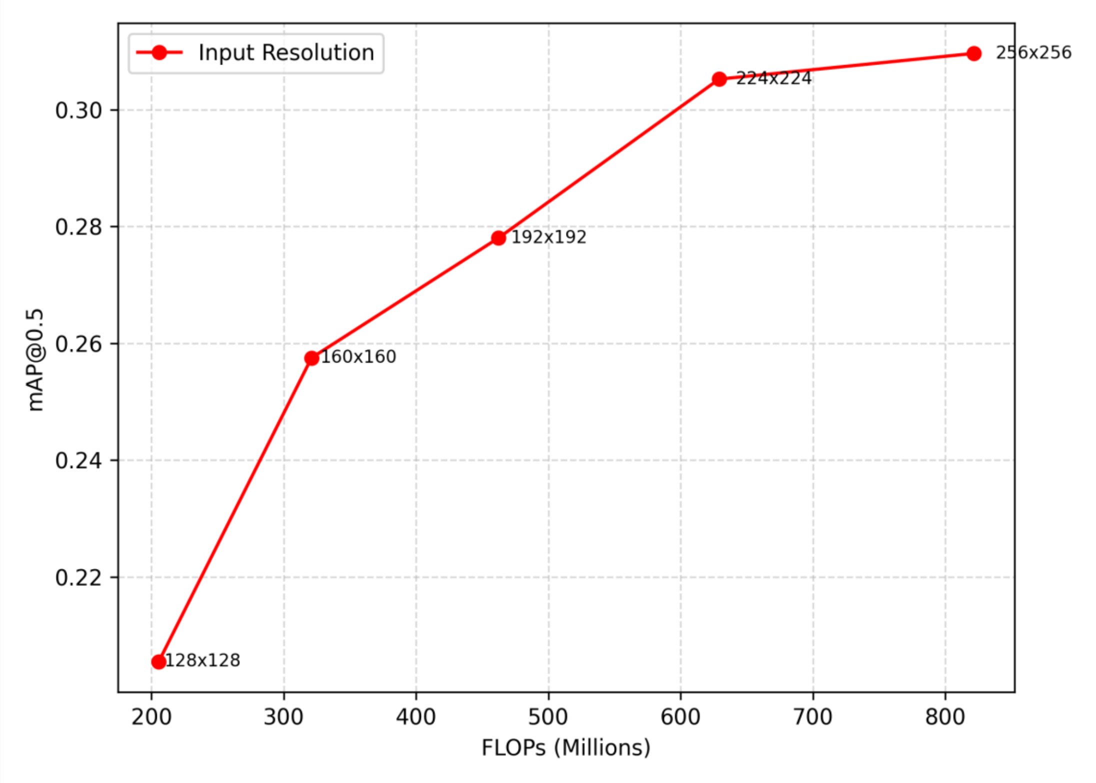
- mAP@0.5 (IoU ≥ 0.5)

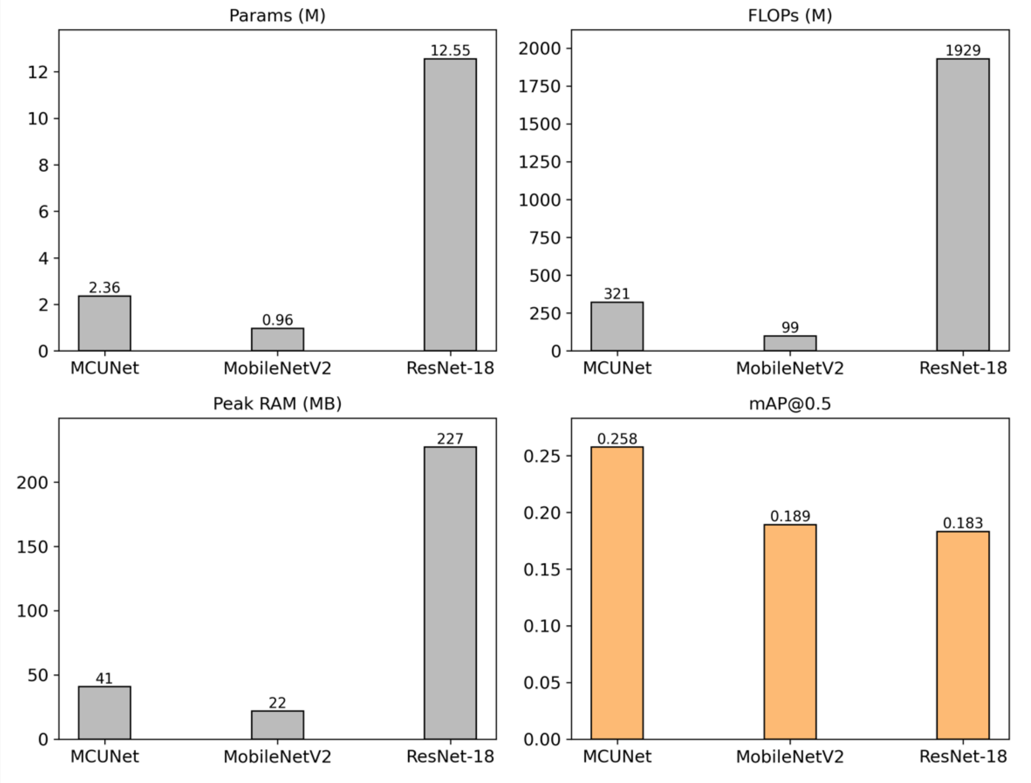
Evaluation

- **Resolution**: Trade-off between input size and FLOPs
- **Backbone**: Compare MCUNet vs mainstream small backbones
- **Quantisation**: Impact of INT8 quantisation on accuracy
- **Deployment**: On-device inference time and stability (100 images, Coral Dev Board Micro, < 1 MB SRAM)

Findings: Pretraining

Resolution Evaluation / Backbone Evaluation





- **Higher resolution gives diminishing boost**:

  1.4× FLOPs → only +0.02 mAP (160 →192)
- **Chosen input size: 160×160** (best trade-off)

- **MCUNet:** highest mAP vs. MobileNetV2 and ResNet-18
- **MobileNetV2:** lowest parameters/FLOPs but weaker accuracy
- **Chosen backbone: MCUNet** (best trade-off)

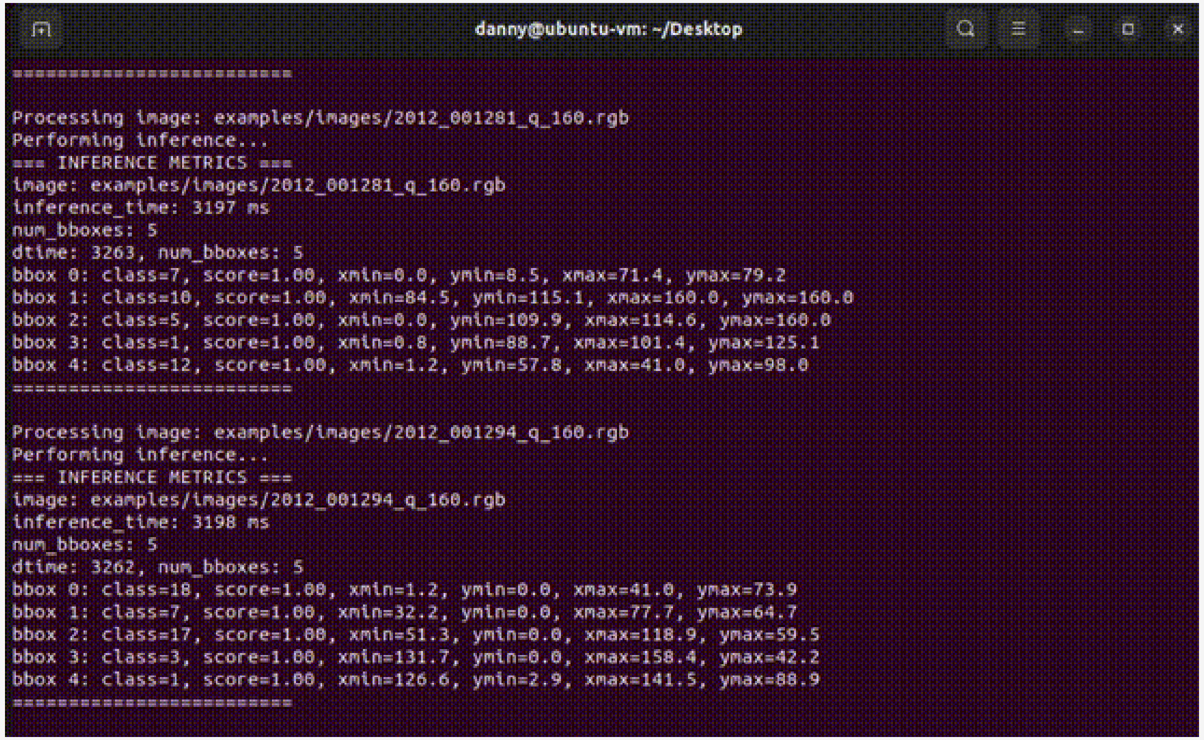
Findings: Pretraining

Quantisation Evaluation / Deployment Evaluation

*Measured on CPU during inference.

Quantisation Scheme	mAP@0.5	Storage (MB)	Peak RAM (MB)*
FP32	0.2575	9.01	15.27
INT8	0.2545	2.61	2.55

- INT8 reduces model size (-71%) and peak RAM (-83%)
- Accuracy remains almost unchanged (−0.003 mAP)
- **INT8: Strong candidate for MCU deployment**



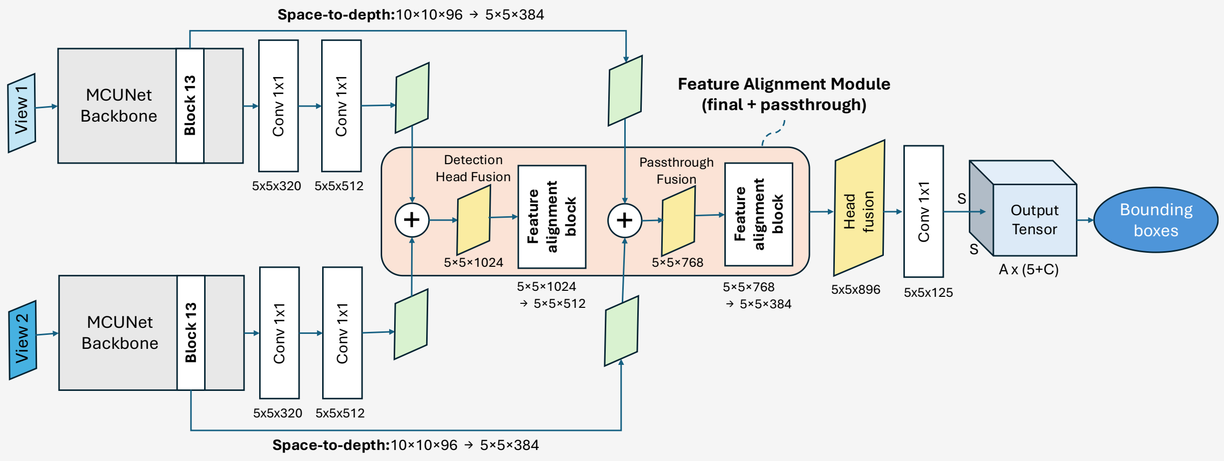
```
=========================

Processing image: examples/images/2012_001281_q_160.rgb
Performing inference...
=== INFERENCE METRICS ===
image: examples/images/2012_001281_q_160.rgb
inference_time: 3197 ms
num_bboxes: 5
dtime: 3263, num_bboxes: 5
bbox 0: class=7, score=1.00, xmin=0.0, ymin=8.5, xmax=71.4, ymax=79.2
bbox 1: class=10, score=1.00, xmin=84.5, ymin=115.1, xmax=160.0, ymax=160.0
bbox 2: class=5, score=1.00, xmin=0.0, ymin=109.9, xmax=114.6, ymax=160.0
bbox 3: class=1, score=1.00, xmin=0.8, ymin=88.7, xmax=101.4, ymax=125.1
bbox 4: class=12, score=1.00, xmin=1.2, ymin=57.8, xmax=41.0, ymax=98.0
=========================

Processing image: examples/images/2012_001294_q_160.rgb
Performing inference...
=== INFERENCE METRICS ===
image: examples/images/2012_001294_q_160.rgb
inference_time: 3198 ms
num_bboxes: 5
dtime: 3262, num_bboxes: 5
bbox 0: class=18, score=1.00, xmin=1.2, ymin=0.0, xmax=41.0, ymax=73.9
bbox 1: class=7, score=1.00, xmin=32.2, ymin=0.0, xmax=77.7, ymax=64.7
bbox 2: class=17, score=1.00, xmin=51.3, ymin=0.0, xmax=118.9, ymax=59.5
bbox 3: class=3, score=1.00, xmin=131.7, ymin=0.0, xmax=158.4, ymax=42.2
bbox 4: class=1, score=1.00, xmin=126.6, ymin=2.9, xmax=141.5, ymax=88.9
=========================
```

- Arena: 759 KB (< 1 MB SRAM limit)
- Latency: 3197 ms avg, 4 ms jitter
- Errors: 0
- **Stable on ultra-low-end MCU**

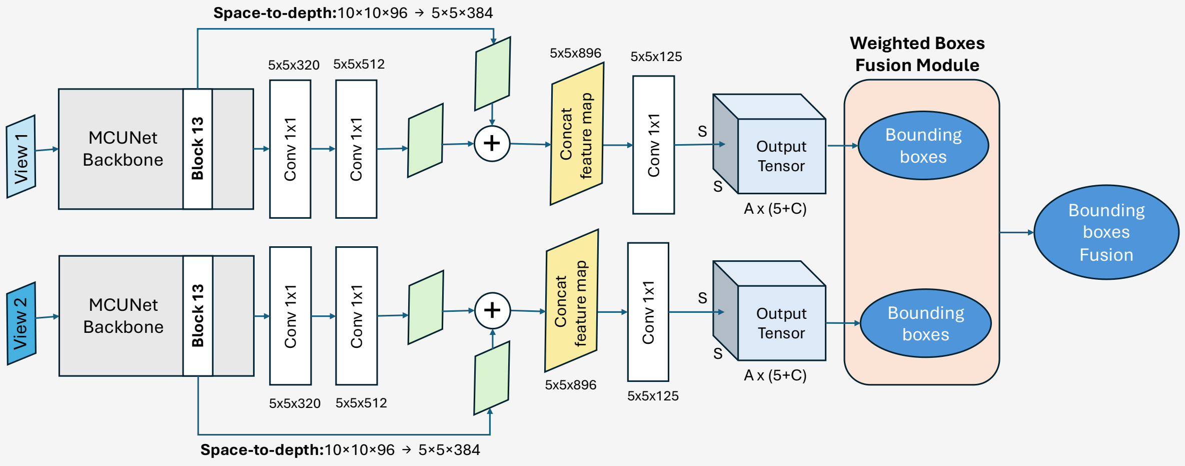
Experiment: Collaborative Inference

Fusion / Dataset / Metric / Evaluation

Feature-Level Fusion: Feature map concatenation



Decision-Level Fusion: Weighted Fusion Boxes



Dataset

- CO3D with CutOut (multi-view, occlusion simulation) under various occlusion combination (30% and 50%)

Metric:

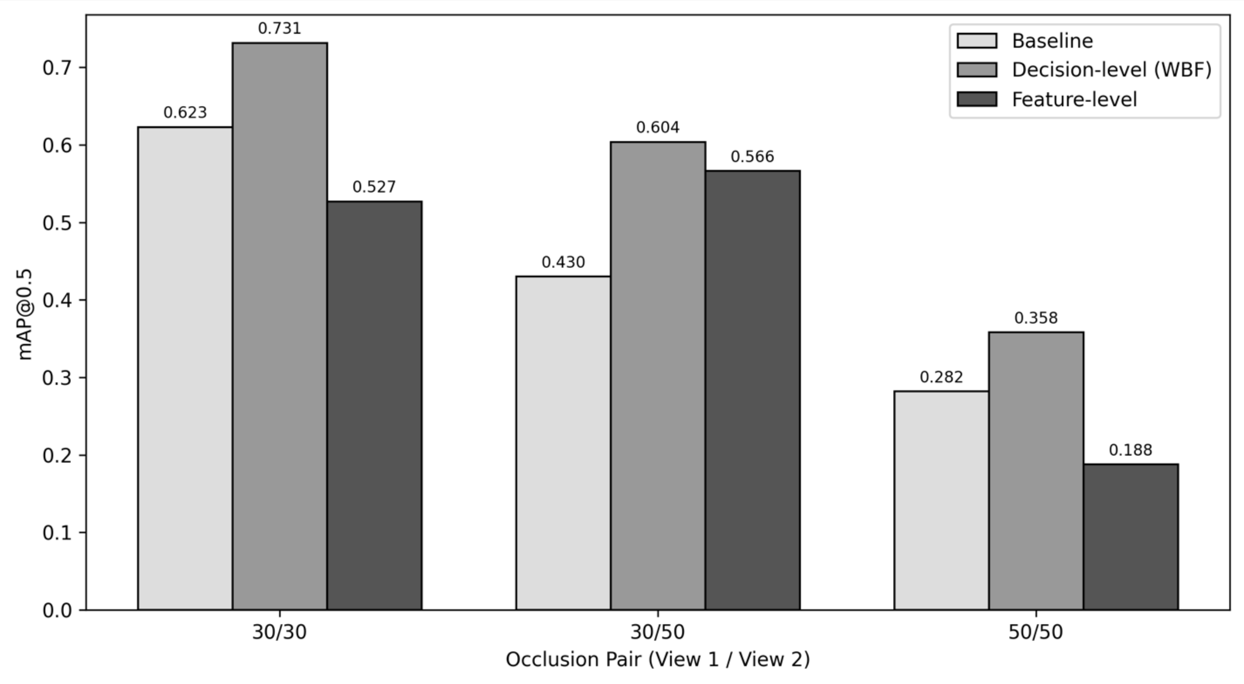
- mAP@0.5 (IoU ≥ 0.5)

Evaluation

- **Fusion Strategies:** Compare feature-level vs. decision-level
- **Scaling:** Accuracy impact of adding a third view
- **Communication:** Trade-off between payload size and accuracy

Findings: Collaborative Inference

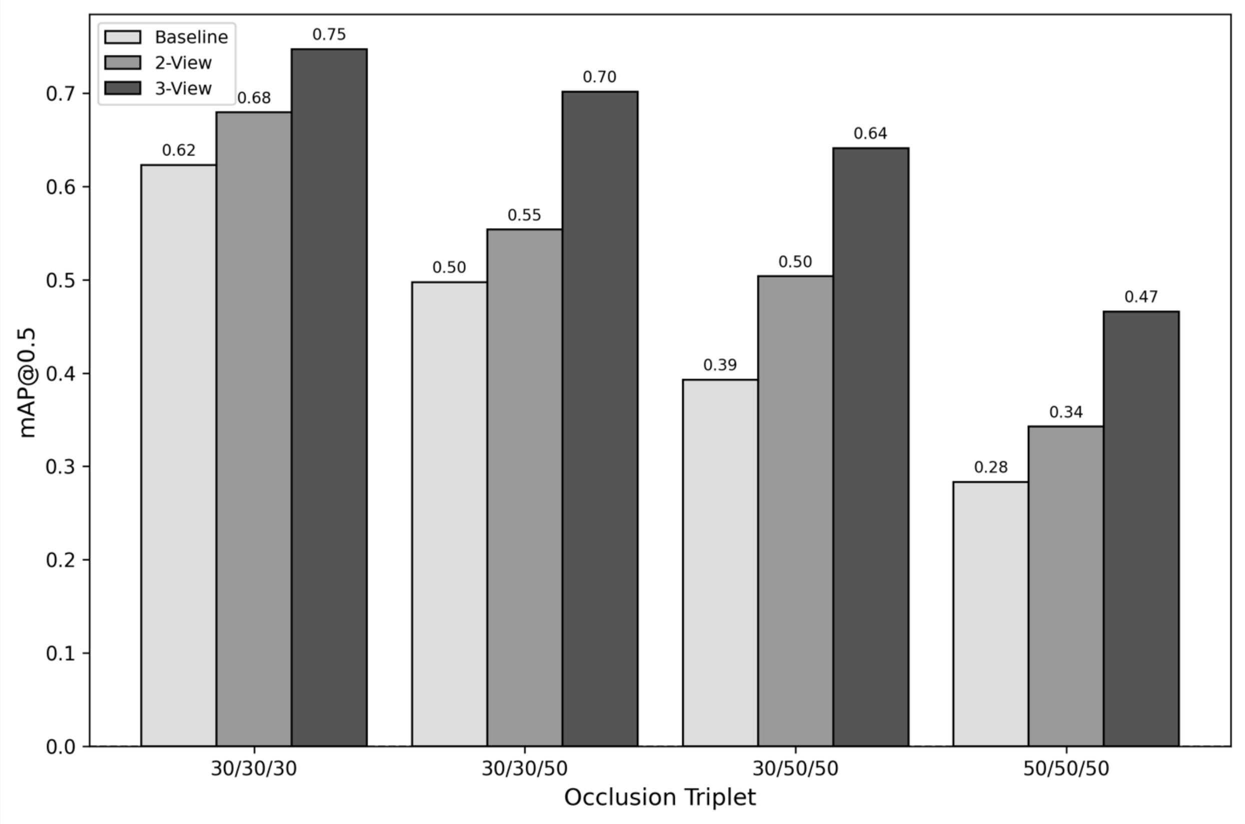
Fusion strategies Evaluation



- **Decision-level fusion (WBF)** outperforms both the single-view baseline and feature-level fusion across all settings
- **Asymmetric occlusion** shows the largest gains for both strategies, leveraging the less-occluded view
- **Feature-level fusion** can underperform baseline (only BatchNorm calibration used)
- **Chosen strategy: Decision-level fusion (WBF)** for scaling experiments

Findings: Collaborative Inference

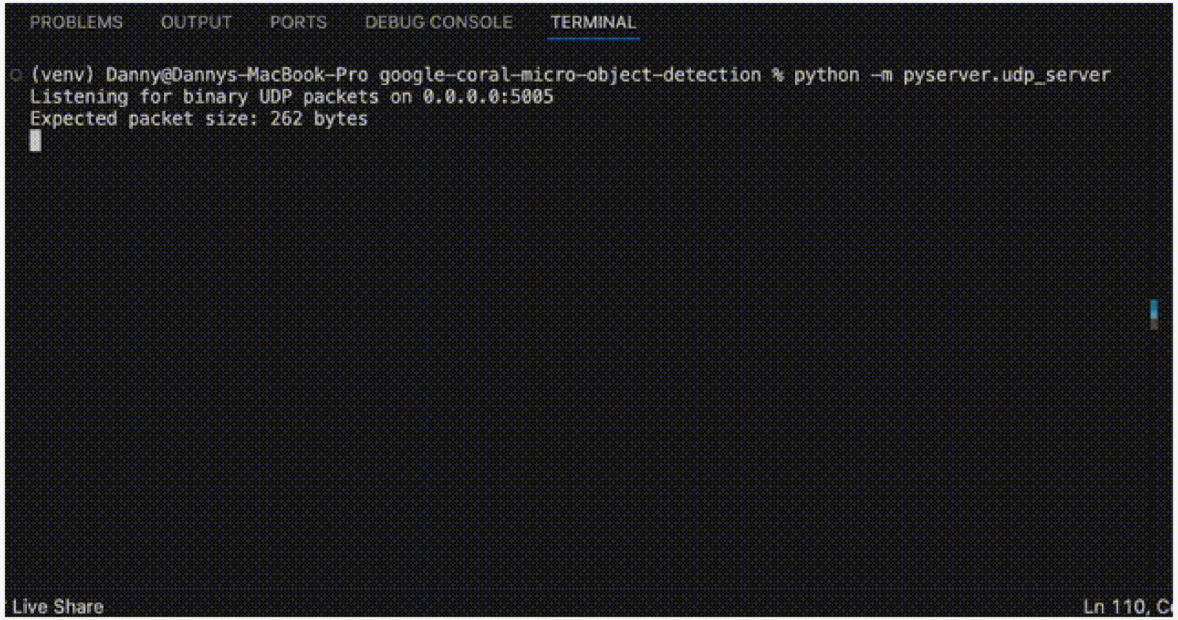
Scaling Evaluation: From One to Three Views



- **Three-view fusion** outperforms both single- and two-view fusion across all settings
- **Occlusion Triplet with heavier occlusion (50%)** shows the largest gains, showing complementary views can recover detections in severe occlusion
- **Trade-off:** Accuracy improves, but communication overhead increases

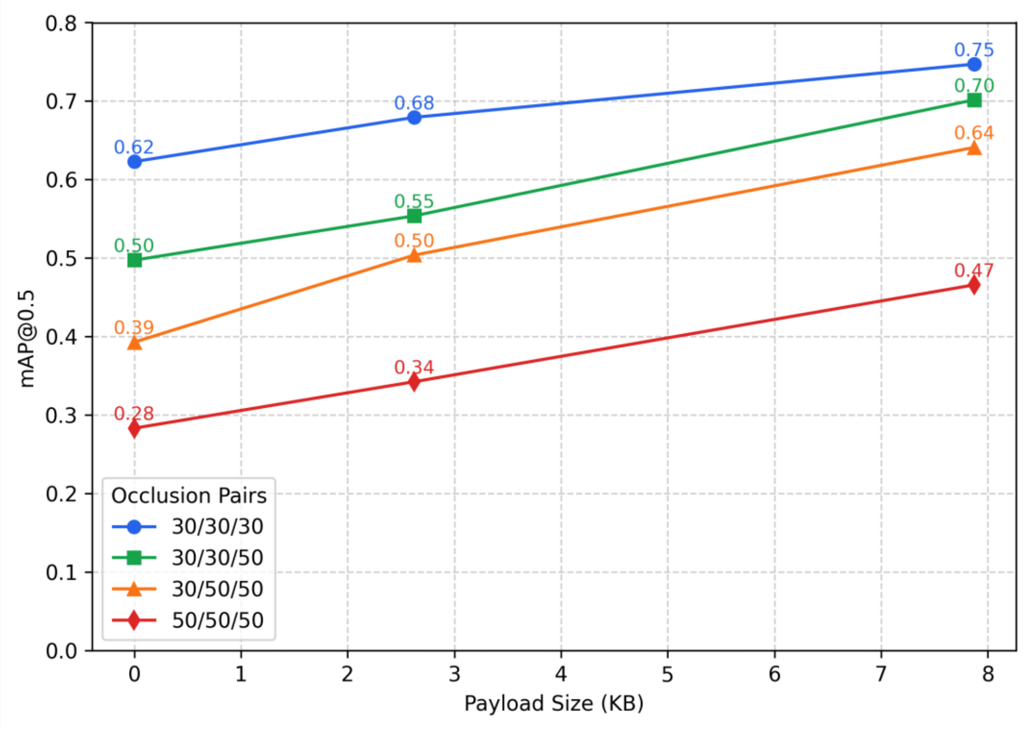
Findings: Collaborative Inference

Communication Evaluation



```
(venv) Danny@Dannys-MacBook-Pro google-coral-micro-object-detection % python -m pyserver.udp_server
Listening for binary UDP packets on 0.0.0.0:5005
Expected packet size: 262 bytes
```

- **80 bounding boxes (1732 bytes)** exceed the Wi-Fi MTU (1500 bytes), causing consistent packet loss
- **60 bounding boxes** is the practical upper bound
- **Chosen payload size:** 60 bounding boxes (1312 bytes)

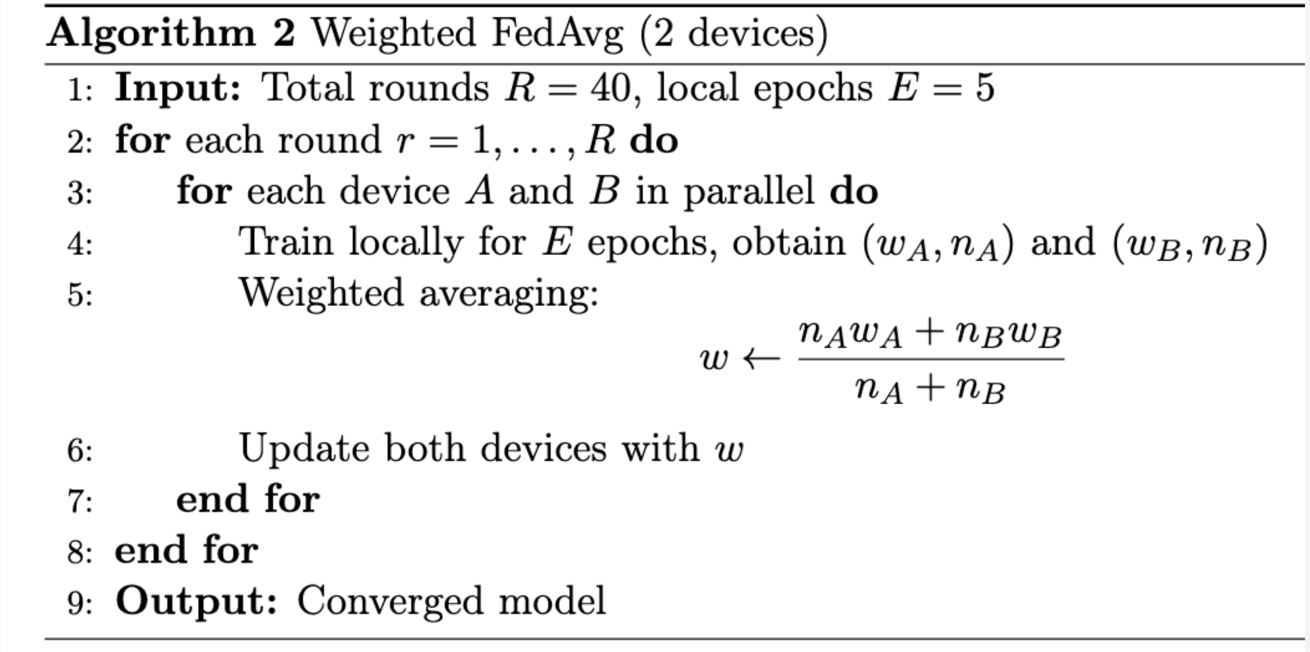


- Three-view fusion improves accuracy by +0.07–0.15 mAP
- Requires **6** information exchanges, **payload reaches ~8 KB**
- **Trade-off:** Accuracy gains vs. device and network communication limits

Experiment: Decentralised Federated Learning

Algorithm / Dataset / Metric / Evaluation

Algorithm



Algorithm 2 Weighted FedAvg (2 devices)

```
**Input:** Total rounds $R = 40$, local epochs $E = 5$
**for** each round $r = 1, \ldots, R$ **do**
    **for** each device $A$ and $B$ in parallel **do**
        Train locally for $E$ epochs, obtain $(w_A, n_A)$ and $(w_B, n_B)$
        Weighted averaging:
            $w \leftarrow \frac{n_A w_A + n_B w_B}{n_A + n_B}$
        Update both devices with $w$
    **end for**
**end for**
**Output:** Converged model
```

Dataset

- CO3D (multi-view)

Metric:

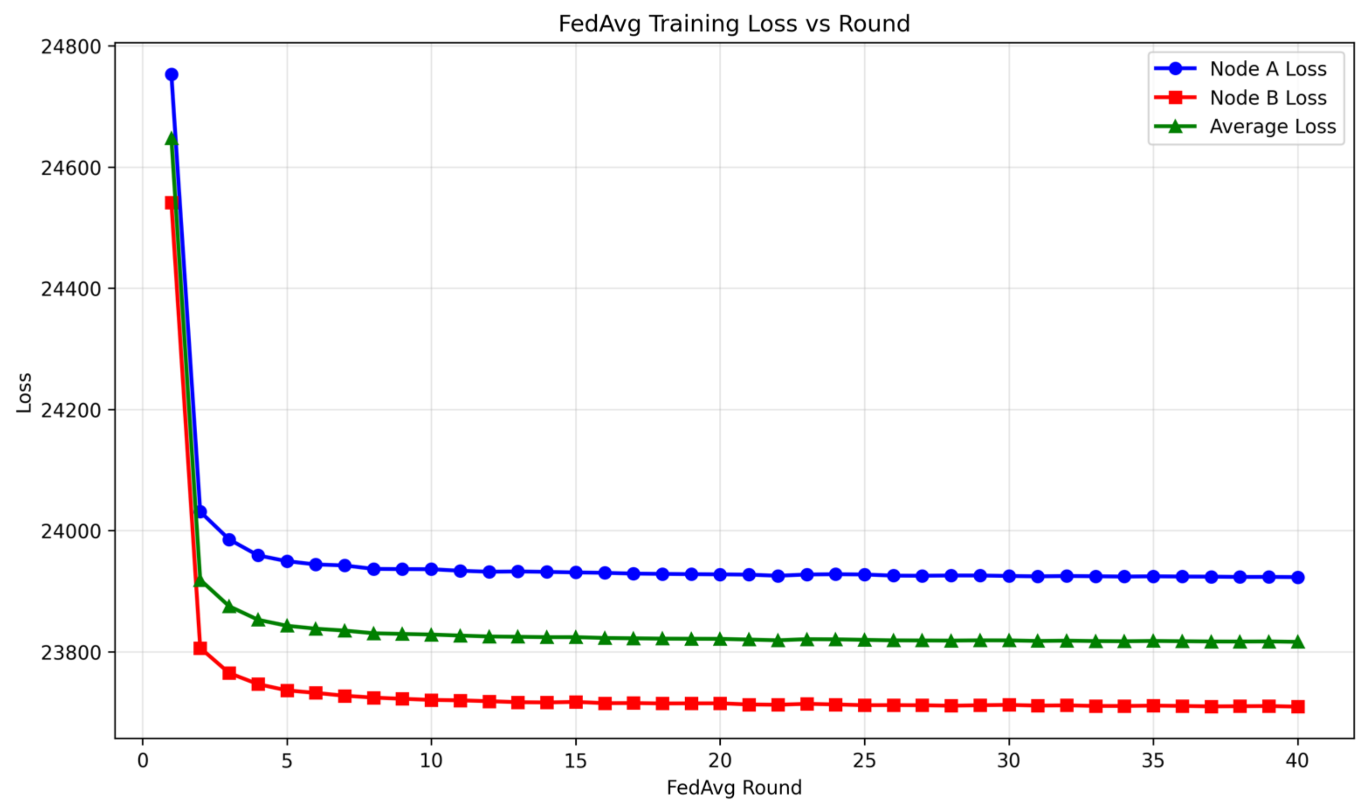
- mAP@0.5 (IoU ≥ 0.5)

Evaluation

- **Convergence:** Check stability of training under non-iid data

Findings: Decentralised Federated Learning

Convergence Evaluation



- **Stable convergence** is observed with FedAvg over 40 rounds and 5 local epochs.
- **High absolute loss (~23,800)** indicates convergence does not yet yield strong performance gains.
- **Limitation**: struggles with non-iid local datasets in decentralised edge scenarios.

Conclusion

- **MCUNet + YOLOv2 with TFLite quantisat**ion enables deployment on ultra-low-end MCUs (<1 MB SRAM).
- **Decision-level fusion (WBF)** consistently outperforms feature-level fusion and the single-view baseline, especially under asymmetric occlusion.
- **Three-view fusion** provides further accuracy gains, but at the cost of higher communication overhead.
- **Weighted Boxes Fusion proves surprisingly effective**. Since it works directly on bounding box outputs, it requires no retraining, no calibration, and no changes to the model structure, yet still delivers the most robust and reliable accuracy gains.
- **Decentralised Federated Learning (DFL)** demonstrates stable convergence without central coordination, but its effectiveness remains limited under non-iid data distributions.

Q & A

IMPERIAL

Thank you

17/09/2025