Edge AI chips for Computer Vision: Applications & Possibilities

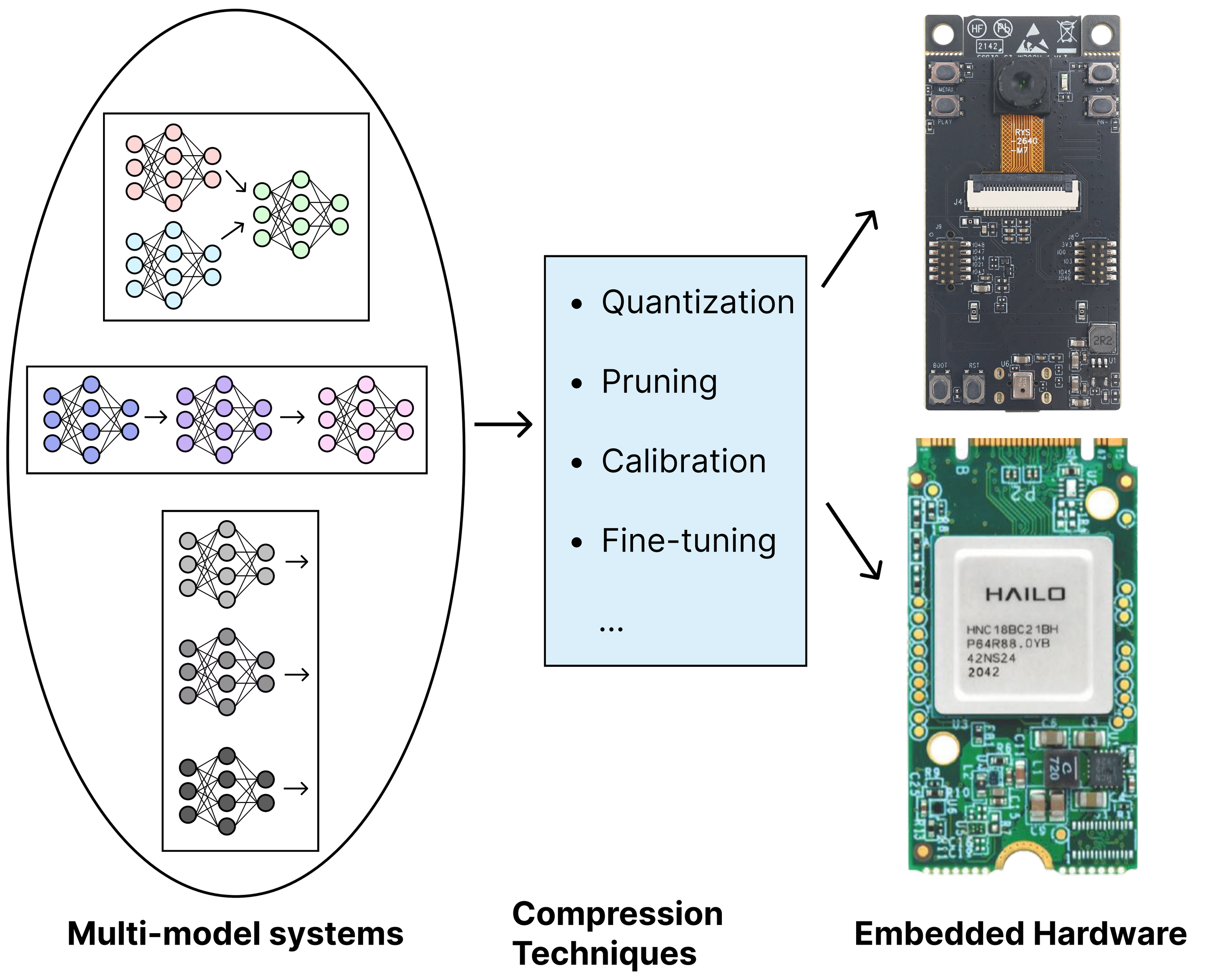
João Prado - Supervisors : Laleh Makarem, Mathieu Salzmann

Goal - One device, multiple models

The deployment of deep learning models for computer vision (CV) at the edge demands a balance between performance, efficiency, and cost. This challenge is heightened when multiple CV models run concurrently on a single edge device.

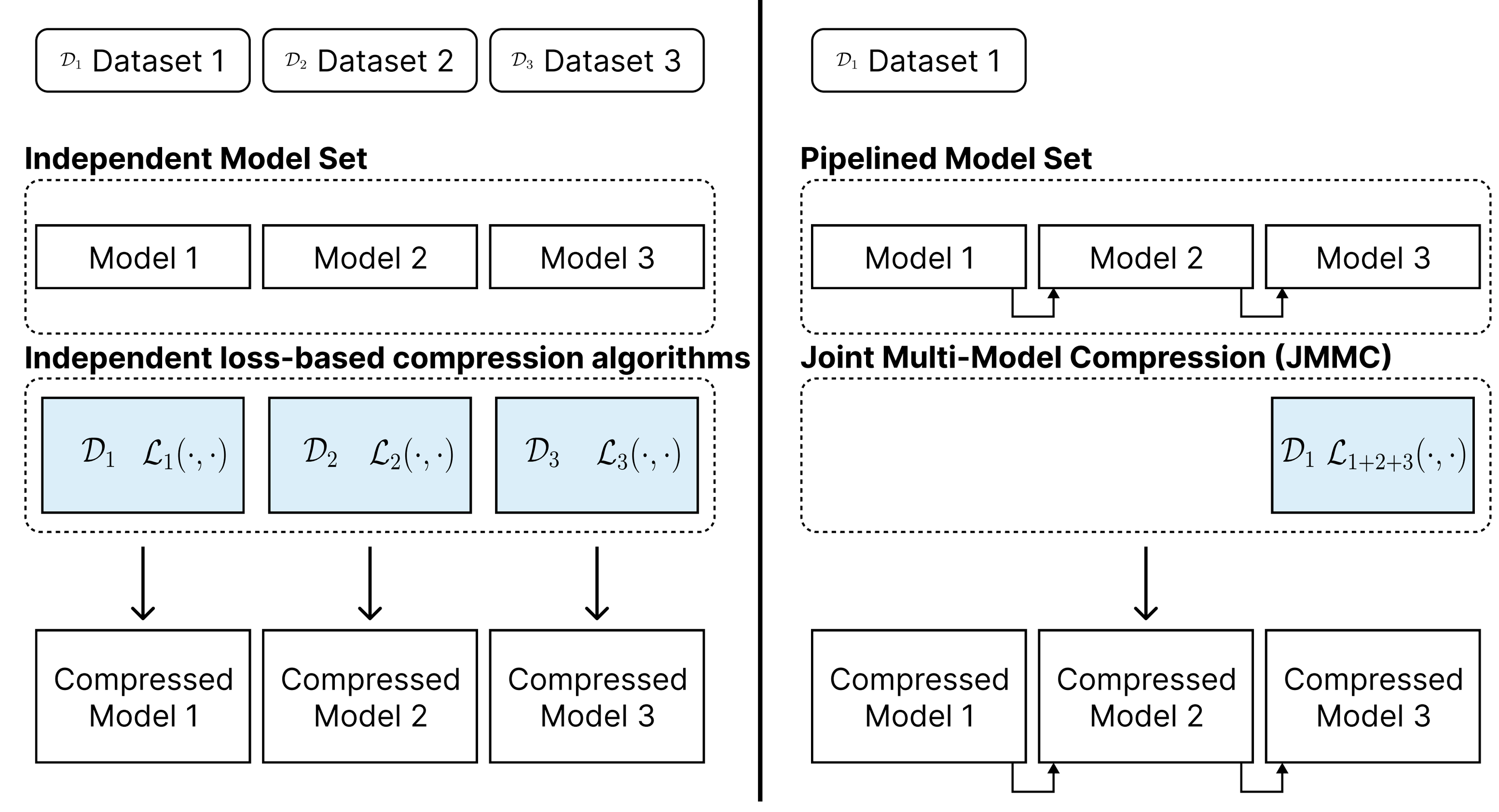
In this project, we benchmark model compression techniques and hardware platforms for deploying several CV models simultaneously in three stages:

1. We consider the problem of detecting human faces in unfavorable imaging conditions as a prototypical CV task requiring the concurrent implementation of multiple image restoration and detection models.
2. We propose **Joint Multi-Model Compression (JMMC)**, an adaptation of Quantization Aware Training (QAT) and pruning techniques in which the multi-model system is fine-tuned as a single unit with an adapted loss function.
3. We port the proposed multi-model system to an edge device containing a Hailo-8 accelerator, we explore the opportunities of parallel inference of the multi-model system and evaluate its efficiency in terms of power consumption and latency.



Proposal: Joint Multi-Model Compression

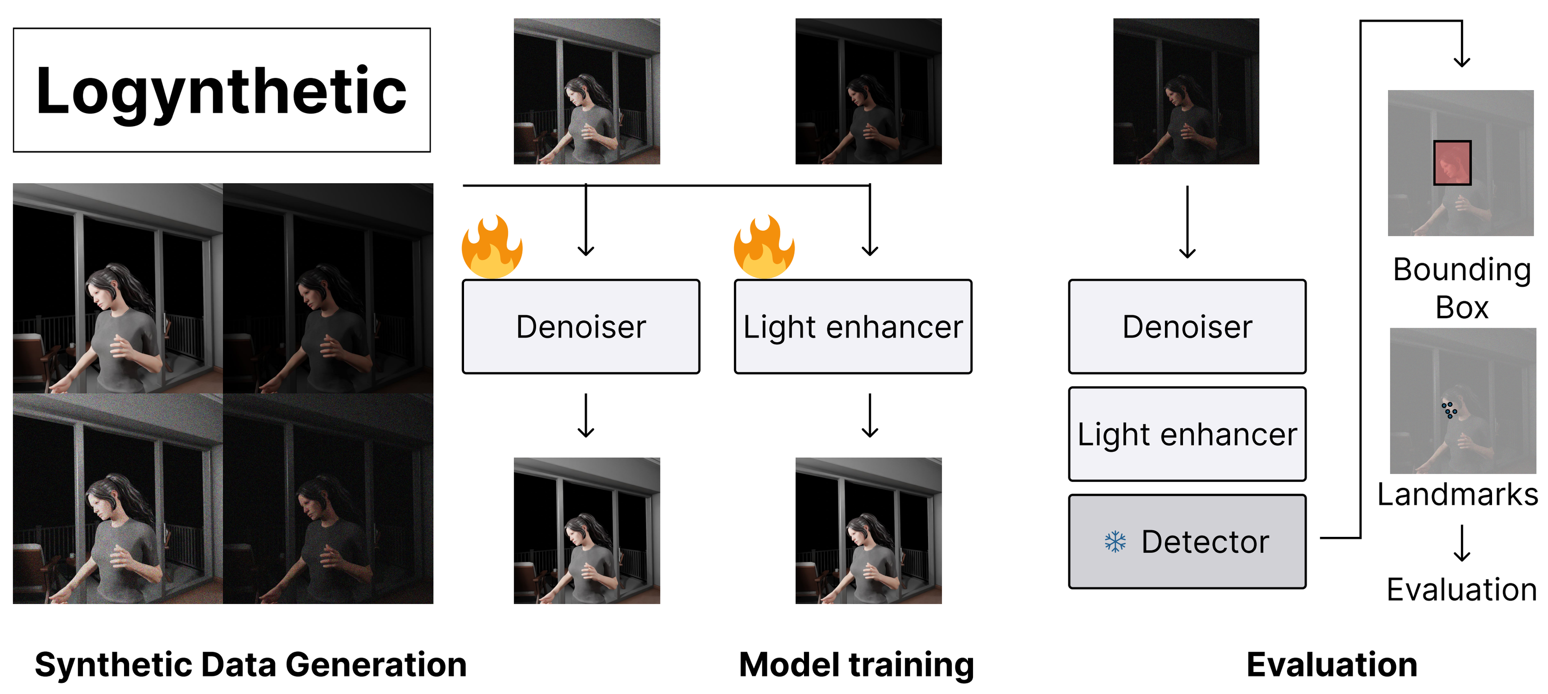
Instead of compressing each model separately with independent datasets and losses, models are **jointly compressed** for a given downstream task, with an adapted loss function.



Case study : Image Enhancement & Detection

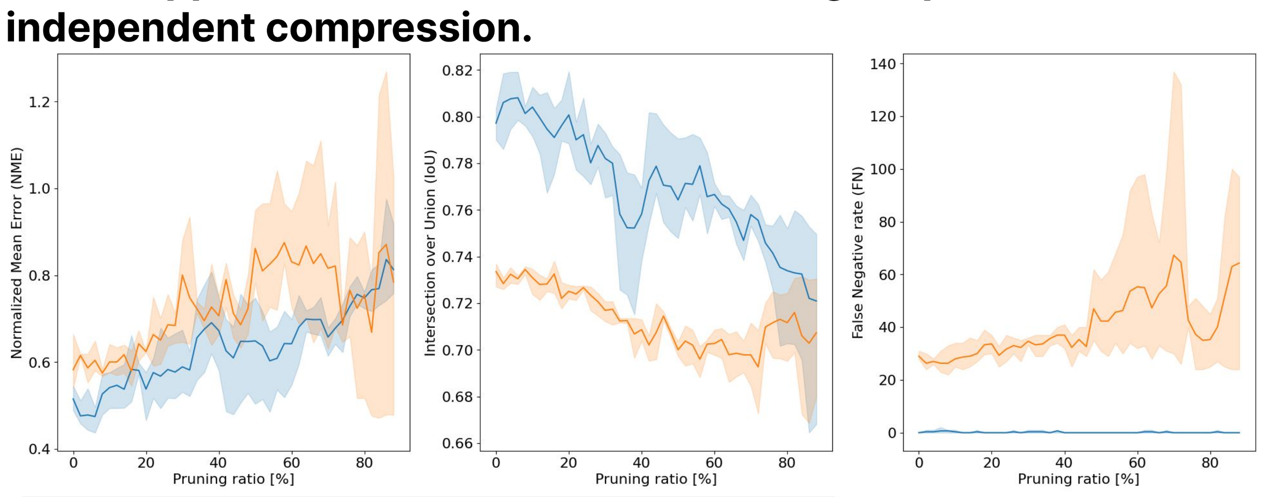
Compression experiments were performed with a proxy task involving **Image Enhancement** and **Detection**. We consider three models:

- Denoiser (PMRID) [1];
- Light-enhancer (PMRID) [1];
- Face and Facial Landmark detector (YuNET) [2];



Results

JMMC applied to Quantization and Pruning outperform independent compression.



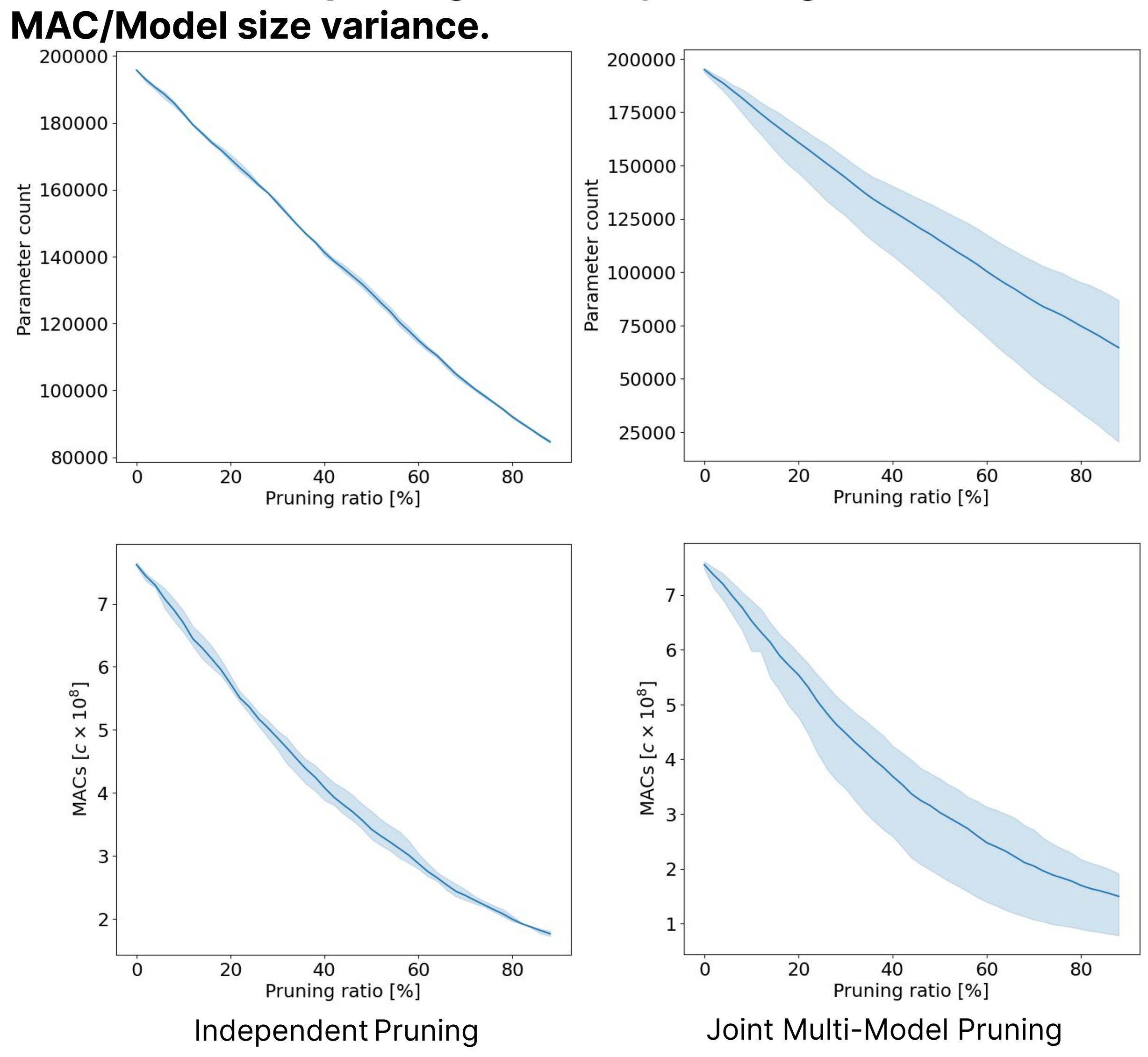
Model Name	BNF	CLE	AdaRound	sQAT	INQ	NME	IoU	FN
Floating Point Baseline						0.579 ± 0.026	0.7339 ± 0.0030	29 ± 1.7
IQ-INQ					✓	2.31 ± 0.17	0.384 ± 0.028	17.67 ± 15.31
IQ-sQAT	✓	✓	✓			0.759 ± 0.081	0.7032 ± 0.0058	46.7 ± 19.4
JMMC-INQ					✓	0.88 ± 0.12	0.716 ± 0.024	**0**
JMMC-sQAT	✓	✓	✓			**0.626 ± 0.023**	**0.7232 ± 0.0080**	38.3 ± 5.859465

Table 3.2: Detection results on degraded images for different quantization configurations of the pipelined system: independently quantized with PTQ and single-shot QAT (IQ-sQAT), independently quantized with INQ (IQ-INQ); jointly quantized with single shot QAT (JMMC-sQAT); jointly quantized with INQ (JMMC-INQ)

Model Name	PSNR	SSIM	NME	IoU	FN
PMRID-denoise	29.92 ± 0.72	0.691 ± 0.030	—	—	—
PMRID-lle	30.3 ± 1.0	0.9614 ± 0.0066	—	—	—
YuNET detector	—	—	—	—	—
IP	—	—	0.78 ± 0.28	0.707 ± 0.026	64 ±37
JMMC	—	—	0.813 ± 0.092	0.721 ± 0.046	0± 0

Table 3.4: (top) Pruning performance for each separate model. (Bottom) Pruning results for different configurations of the pipelined system: independently pruned models (IP); jointly pruned (JMMC); Results correspond to a pruning ratio of 90%.

JMMC increases pruning efficiency, although with increased MAC/Model size variance.



Future Outlook

Venues of future work:

- Adapting JMMC to hardware-aware compression pipelines that adapt to a given computing and memory budget.
- Benchmarking closed-source compression pipelines across different providers.
- Implementing JMMC to other downstream tasks and platforms, targeting other use cases of multi-systems in CV.

References

[1] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. *Practical Deep Raw Image Denoising on Mobile Devices.* 2020. arXiv
[2] Wei Wu, Hanyang Peng, and Shiqi Yu. YuNet: *A Tiny Millisecond-level Face Detector.* Apr. 2023. DOI: 10.1007/s11633-023-1423-y. U R L: http://dx.doi.org/10.1007/s11633-023- 1423-y.