

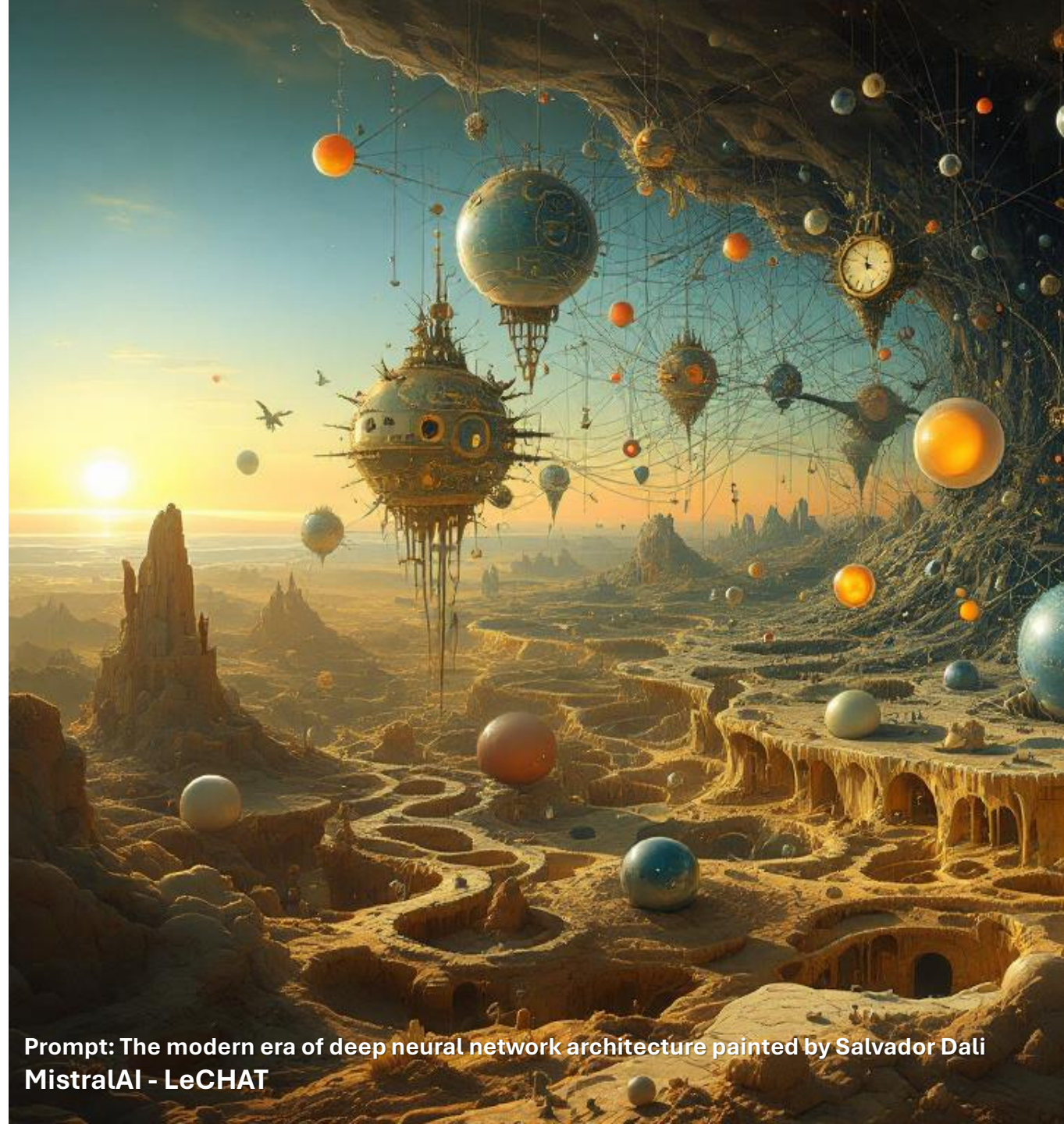
LeYOLO, nouvelle architecture pour la détection d'objets sur appareils embarqué

Lilian Hollard, Lucas Mohimont, Nathalie Gaveau, Luiz-Angelo Steffenel

Doctorant, LICIIS / RIBP



PFIA - APIA



Prompt: The modern era of deep neural network architecture painted by Salvador Dali
MistralAI - LeCHAT

Introduction & Motivations

Recherche sur l'apprentissage profond

1. Thèse
2. Optimisation des modèles

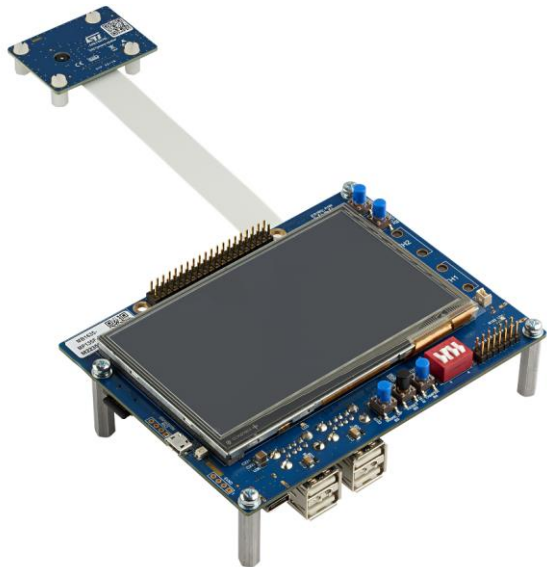
IA Appliquée

1. **Viticulture**
2. Météorologie
3. Covid



Motivations et problématique

- Recherche sur l'apprentissage profond
 - Thèse
 - Optimisation des modèles

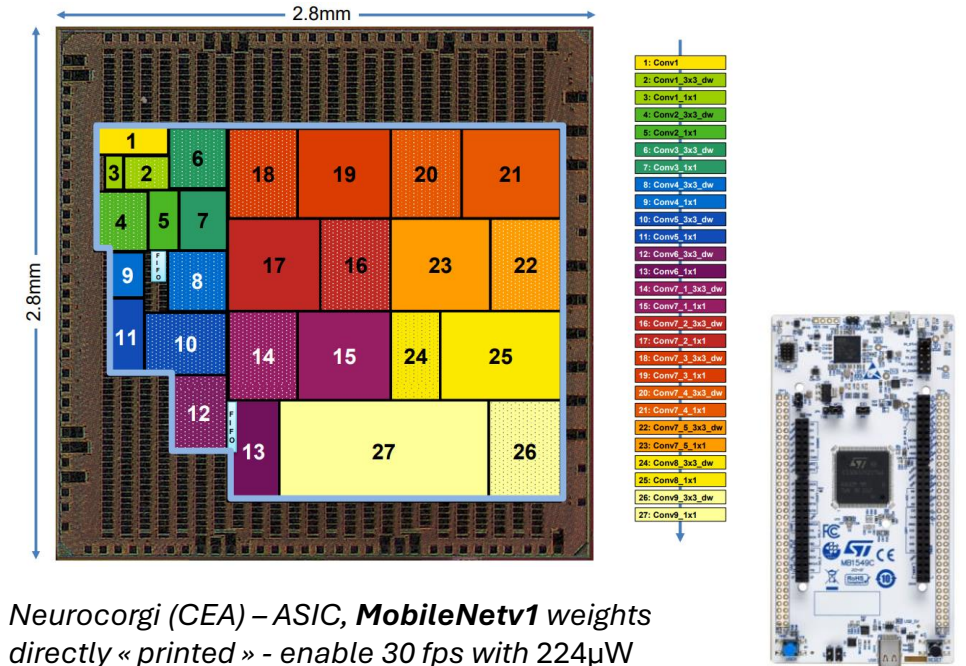


STM32MP257FAI3

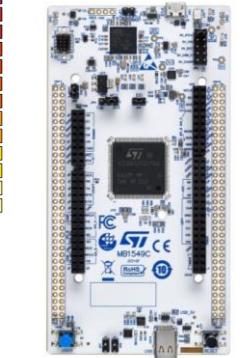
YOLOv8 Nano : **1235** ms

YOLOv9 Tiny : **1371** ms

YOLOv10 Nano : **1099** ms



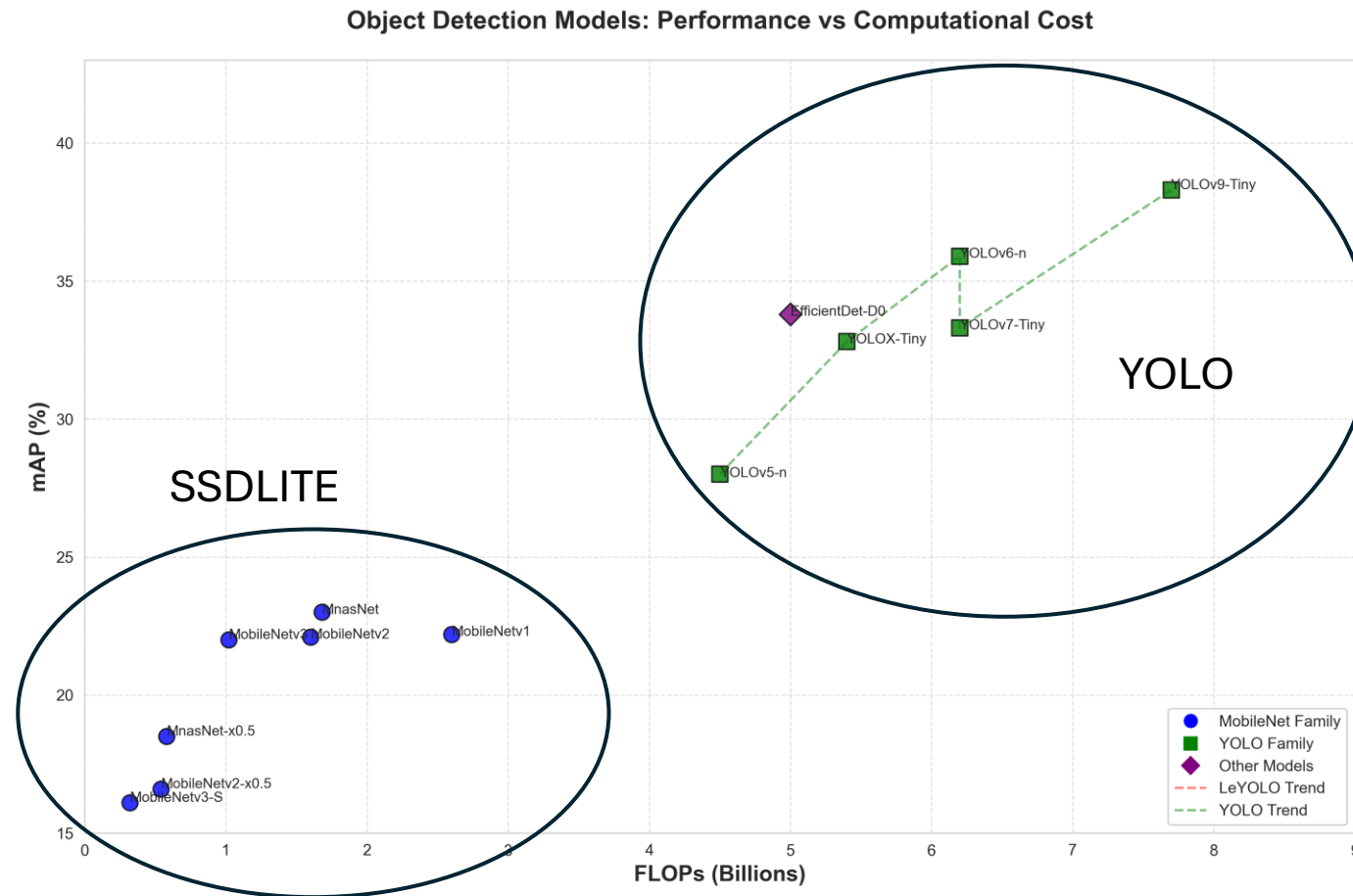
Neurocorgi (CEA) – ASIC, **MobileNetv1** weights directly « printed » - enable 30 fps with 224μW



Nucleo-144
(STMicroelectronics)
– **MobileNetv1**

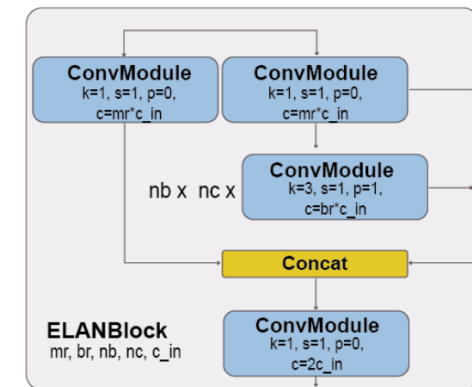
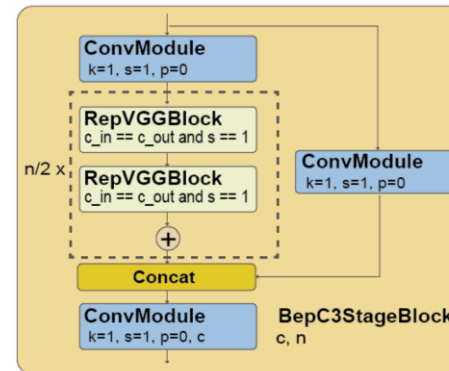
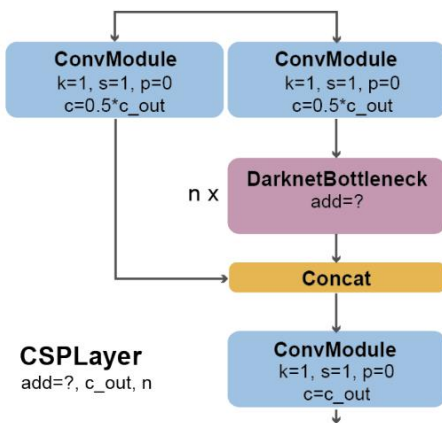
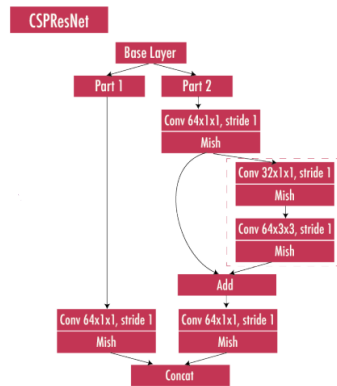
Objectifs

- Réduire le coût de calcul des détecteurs d'objets en s'intéressant à l'architecture du réseau de neurones.

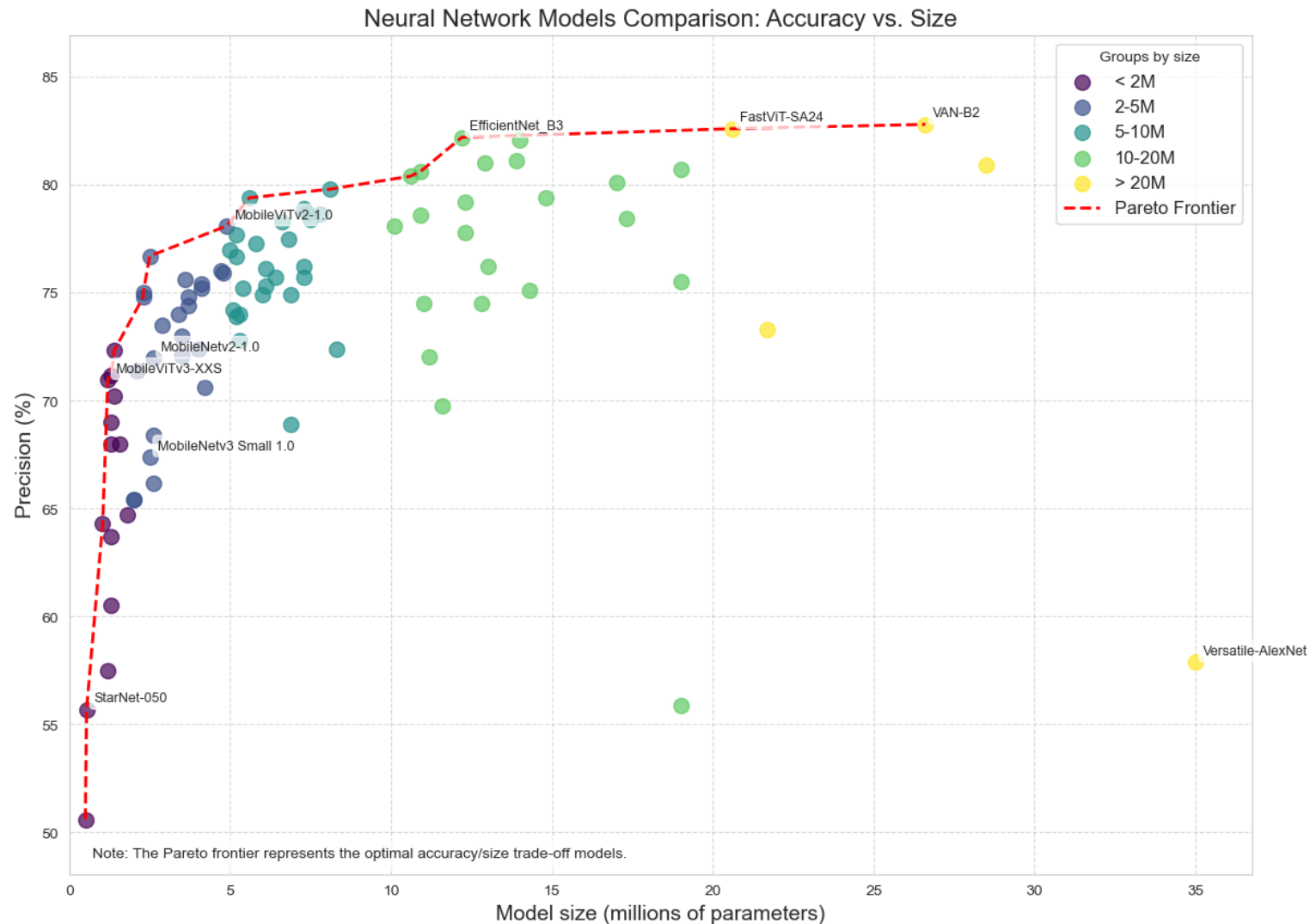


Méthodologie – Comprendre les coûts

- YOLOv4 à v10
 - Utilise des blocs multibranches
 - Rapide, mais utilise beaucoup de paramètres et cout de calcul théorique (FLOP)



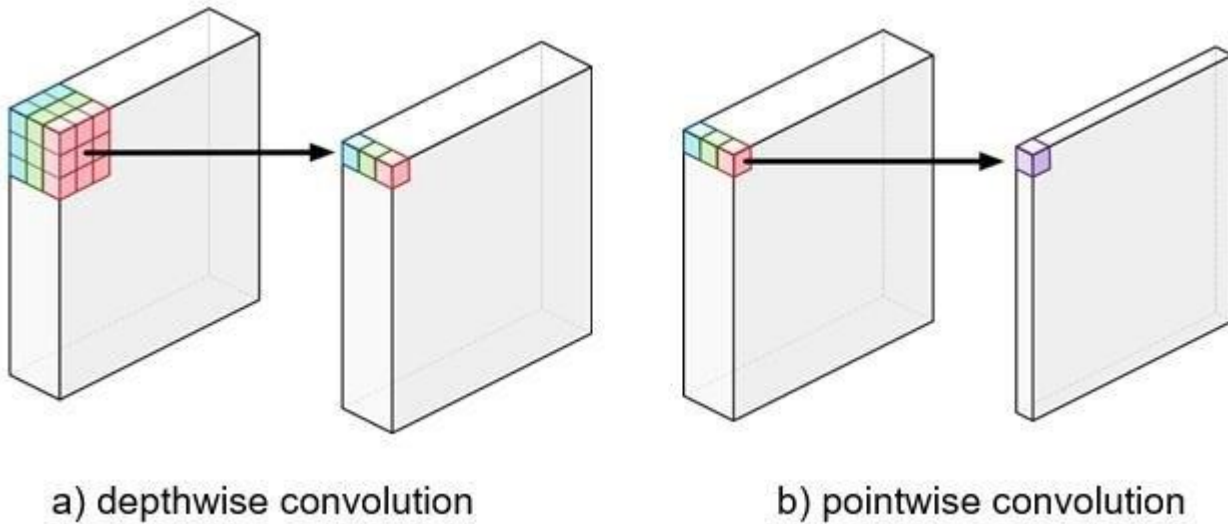
Méthodologie – Observer la précision



Avant LeYOLO – **Aucun** modèle YOLO n'utilisait les inverted bottleneck

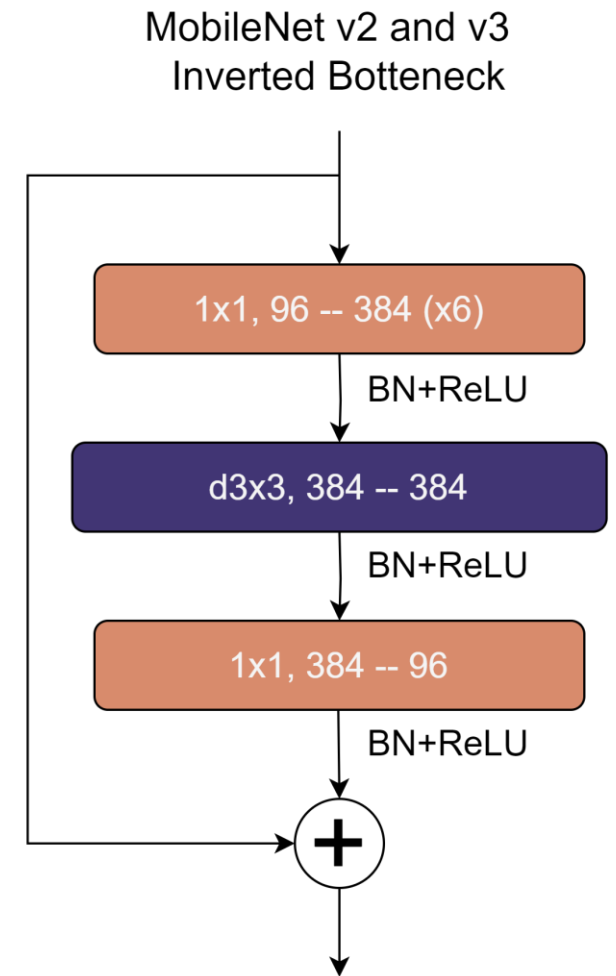
Méthodologie – Inverted Bottleneck

- Des classifieurs aux détecteurs d'objets
 - SOTA : Inverted Bottleneck

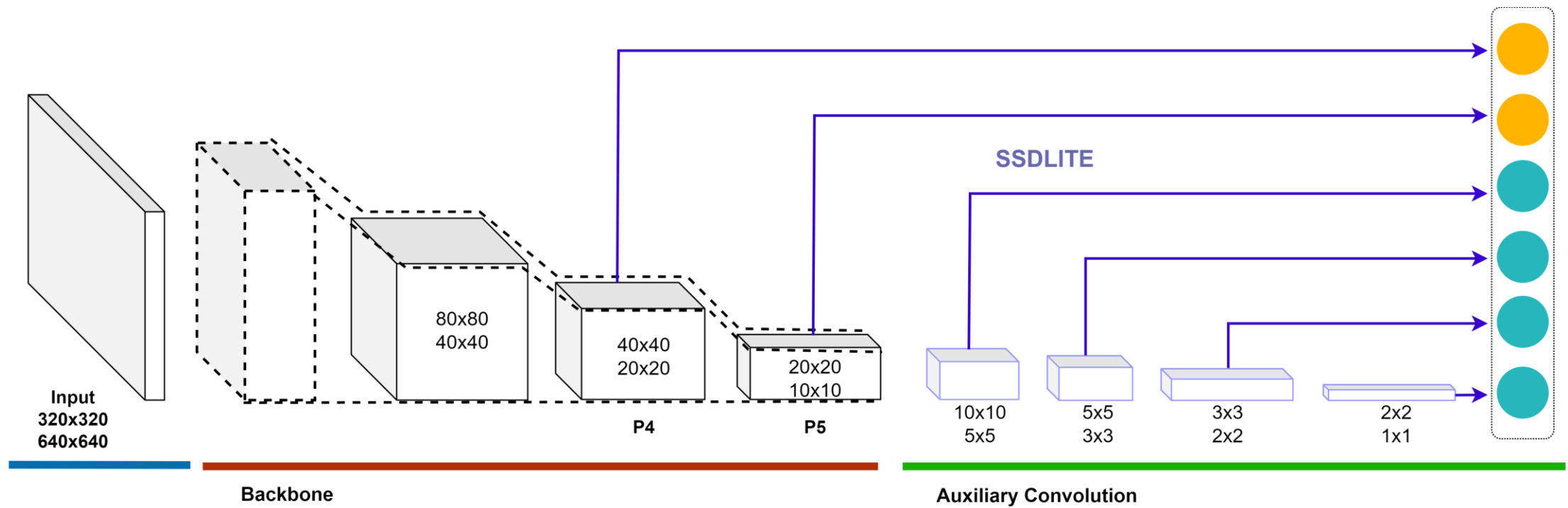


Convolution standard : $3 \times 3 \times 256 \times 256 = 589\,824$ paramètres

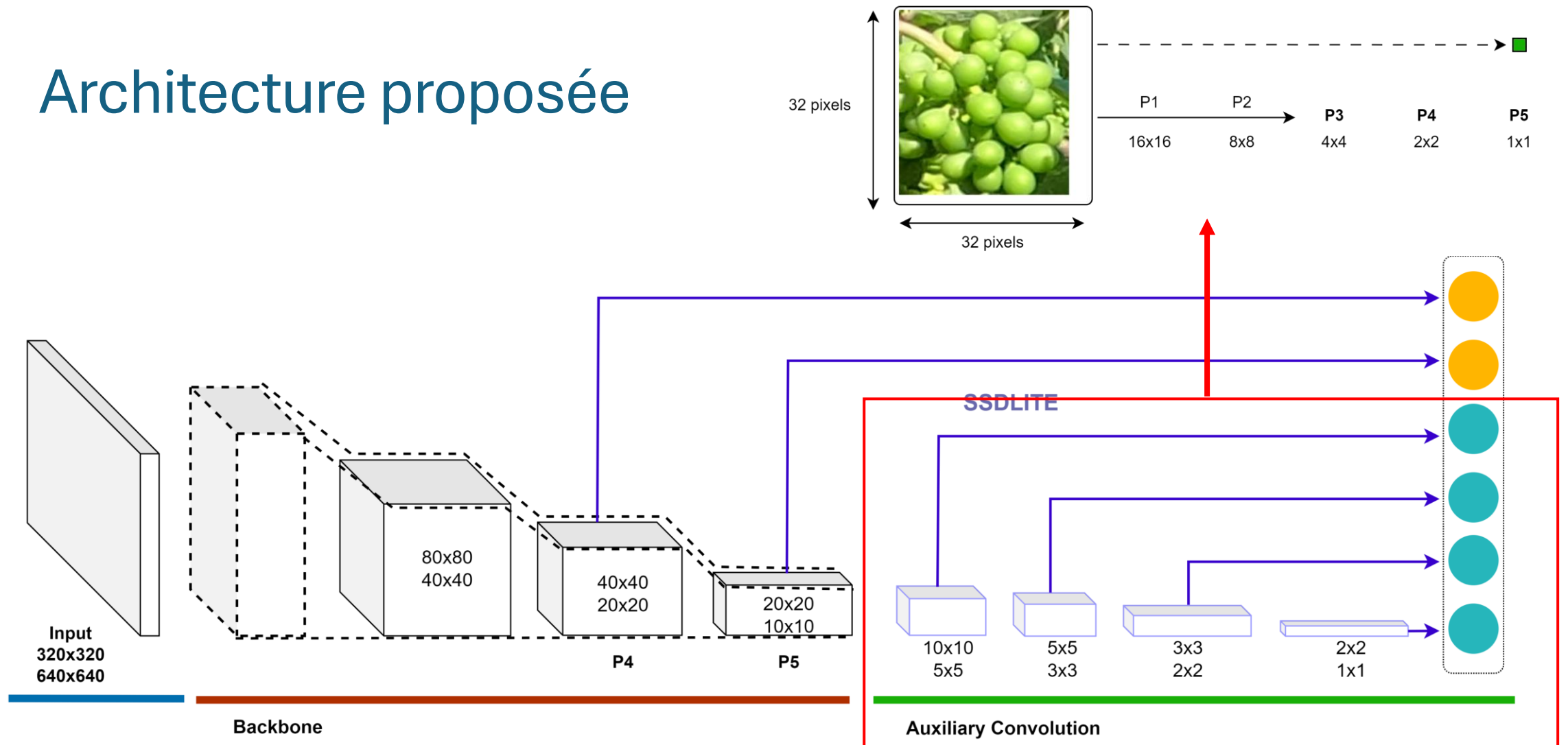
Depthwise + pointwise : $3 \times 3 \times 1 \times 256 + 1 \times 1 \times 256 \times 256 = 67\,840$ paramètres



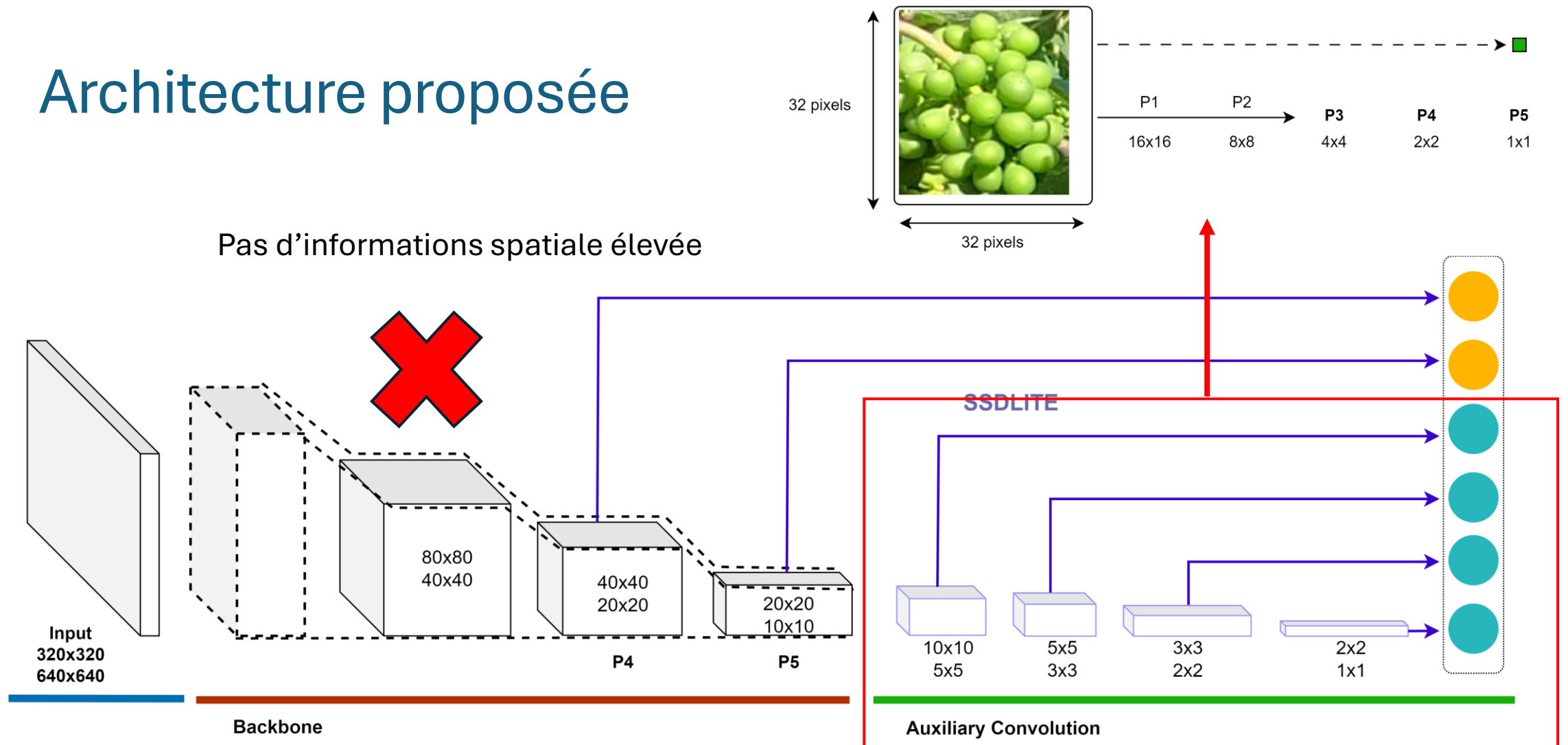
Architecture proposée



Architecture proposée



Architecture proposée



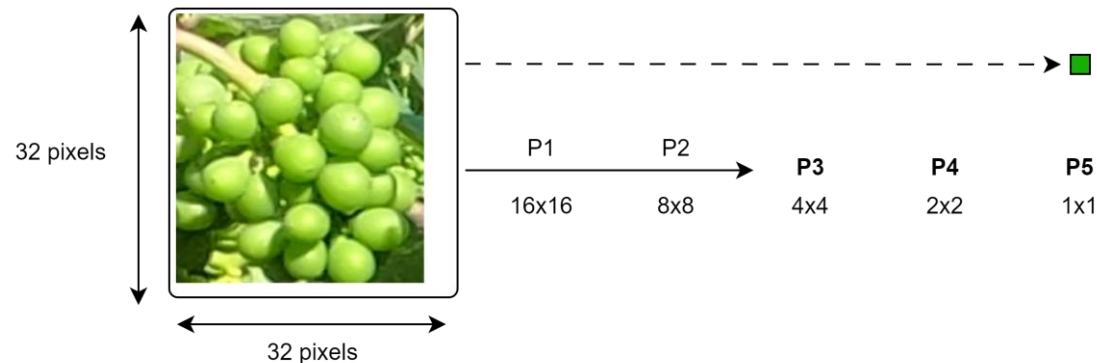
FLOP Théorique : $3 \times 3 \times 64 \times 64 \times 80 \times 80 = 235$ MFLOP

FLOP Théorique : $3 \times 3 \times 64 \times 64 \times 40 \times 40 = 58$ MFLOP

Convolution auxiliaire à trop forte
sémantique

Architecture proposée

- S'inspirer de ce qui se fait du côté des YOLO
 - Partager les informations sémantiques
 - Utiliser **P3**
- Retirer les convolutions auxiliaires, trop peu d'intérêts concrets
 - + de 40% de nos grappes font moins de 30x30p
 - Auxiliaire : - de 1x1 pixel d'information

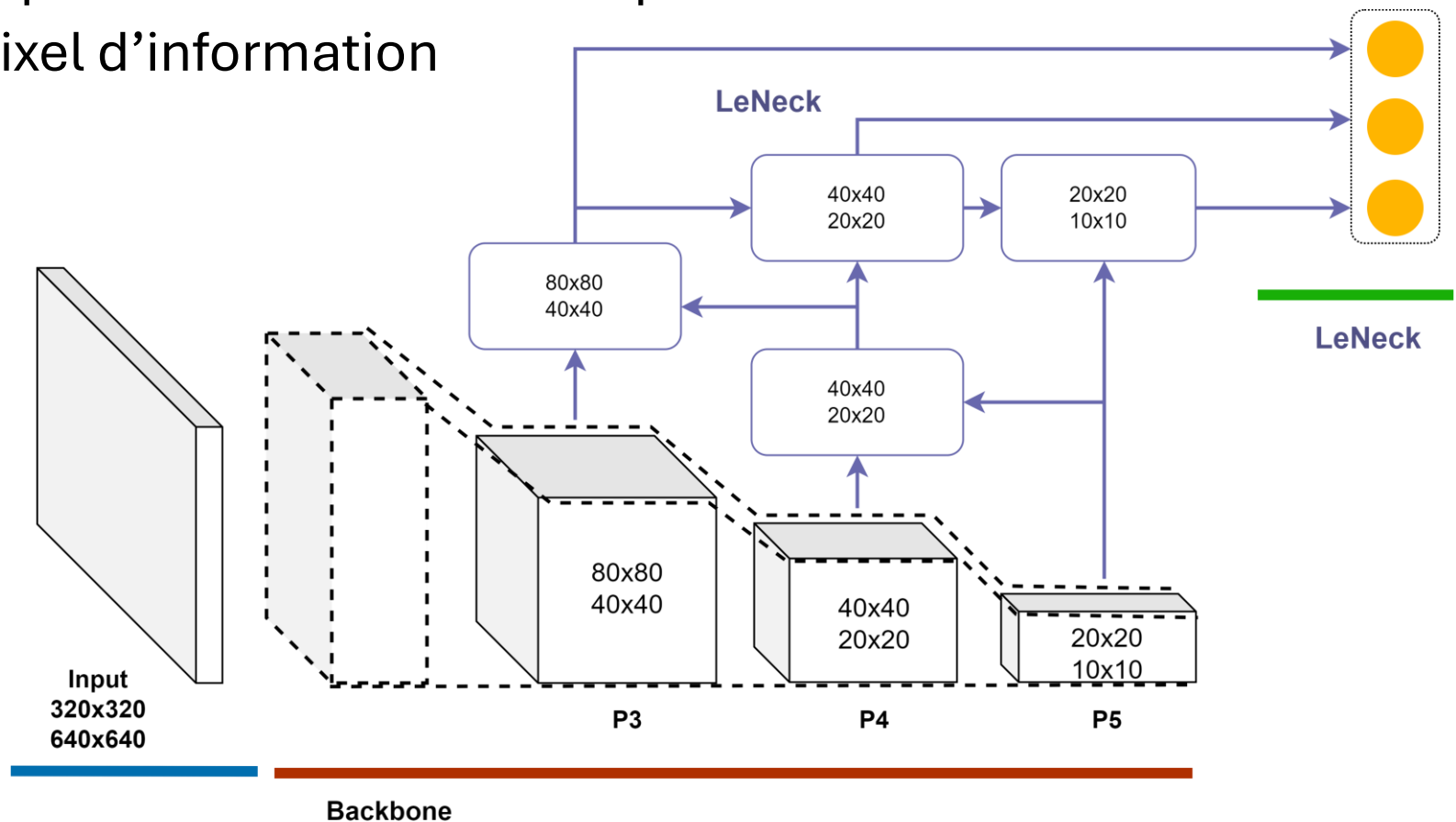


Architecture proposée – LeNeck

- Retirer les convolutions auxiliaires, trop peu d'intérêts concrets
 - + de 40% de nos grappes font moins de 30x30p
 - Auxiliaire : - de 1x1 pixel d'information

Gains:

- Moins de boîtes de détection (6 niveaux vs 3)
- Coût + élevé à 80x80 et 20x20
80x80 – taille spatiale
20x20 – trop de couches
- Optimisation des coûts en utilisant **uniquement P4** comme agrégateur d'information sémantique



Résultats – MSCOCO

Vitesse moyenne avec :

STM32MP257FAI3

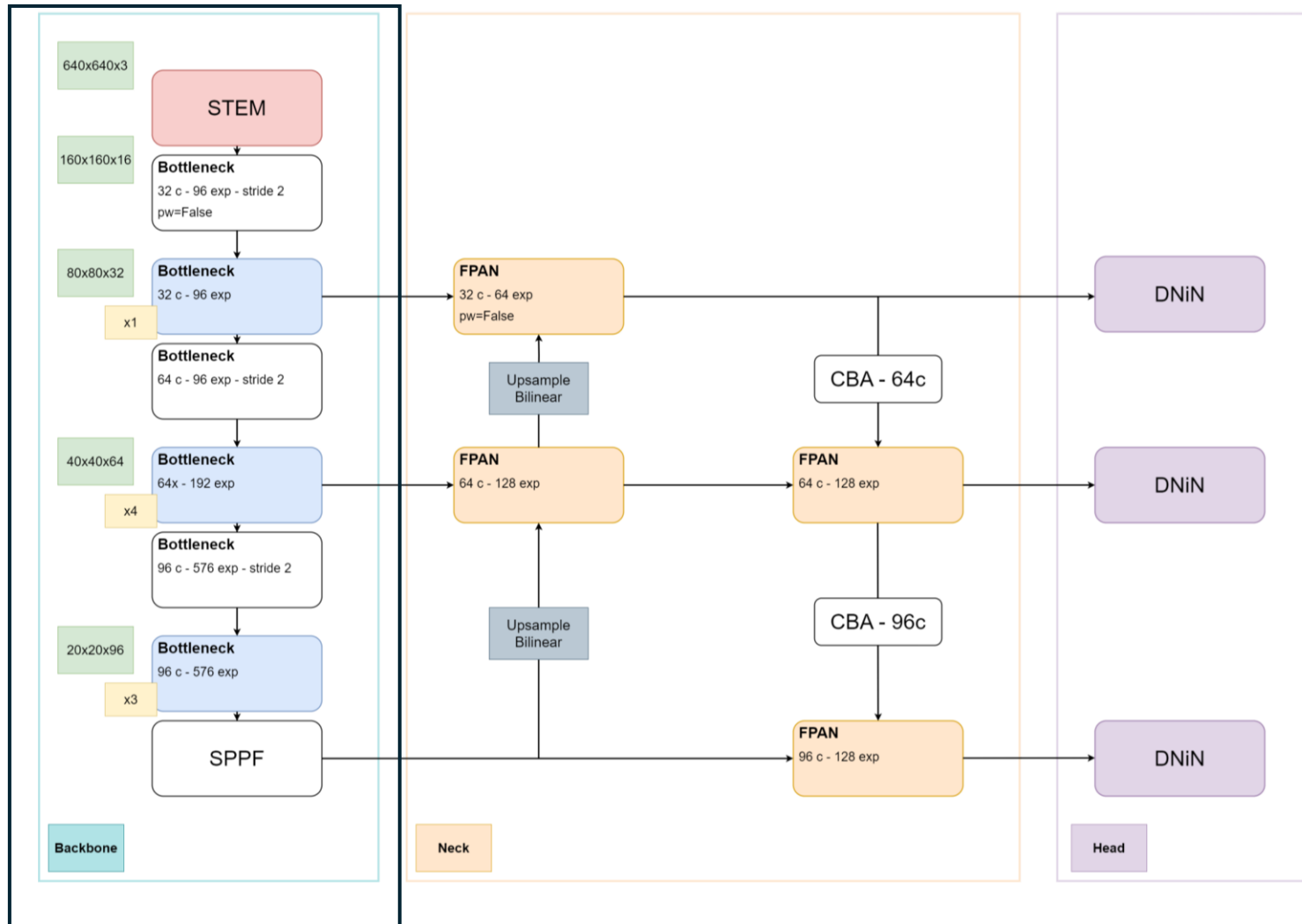
STM32N6570-DK

+28% mAP

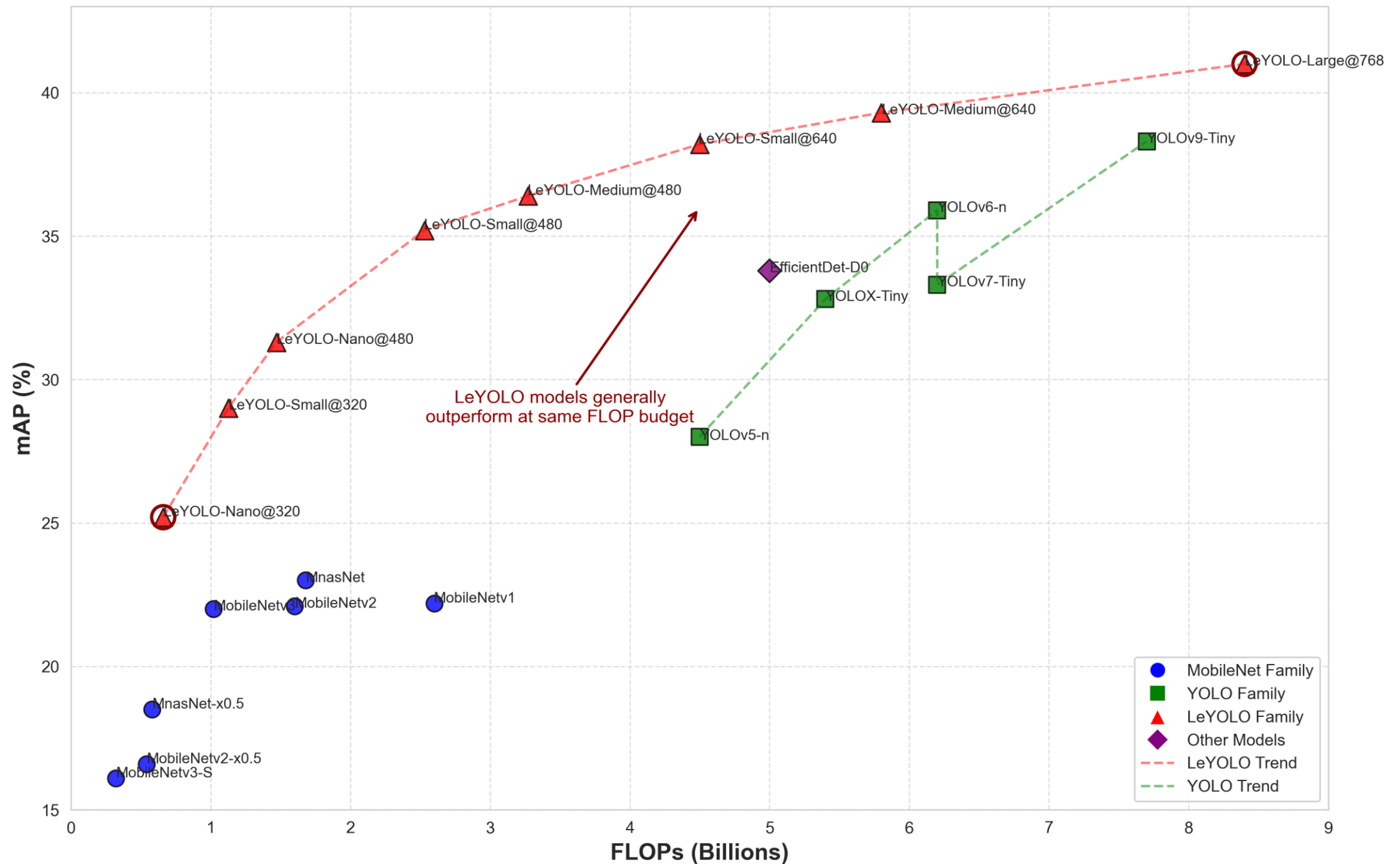
Models	SSDLite	LeNeck	SSDLite	LeNeck
	Speed(ms)		mAP.95	
V3-Small	46.91	50.19	16.0	21.3
V3-Large	121.6	129.2	22	28.1
V2-1.0	104	103.5	22.1	28.6
MNASNet 0.5	86.31	36.03	18.5	24.6
MNASNet	161.3	53.78	23	28.9

Models	SSDLite	LeNeck	SSDLite	LeNeck
	Parameters(M)		mAP.95	
V3-Small	2.49	1.34	16.0	21.3
V3-Large	4.97	3.33	22	28.1
EfficientDetD0	3.9	3.29	34.6	37.1
V2-0.5	1.54	0.98	16.6	23.3
V2-1.0	4.3	2.39	22.1	28.6
MNASNet 0.35	1.02	0.7	15.6	20.0
MNASNet 0.5	1.68	1.22	18.5	24.6
MNASNet	4.68	2.8	23	28.9

LeYOLO – YOLO générique

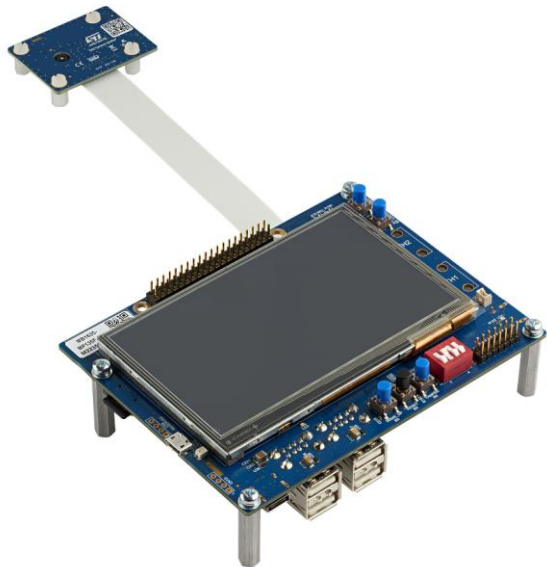


Object Detection Models: Performance vs Computational Cost



Résultats

- LeYOLO Small est aussi précis que YOLOv8, v9 et v10
- Plus rapide : **-26%** de temps d'inférence



STM32MP257FAI3

LeYOLO Nano : **596** ms

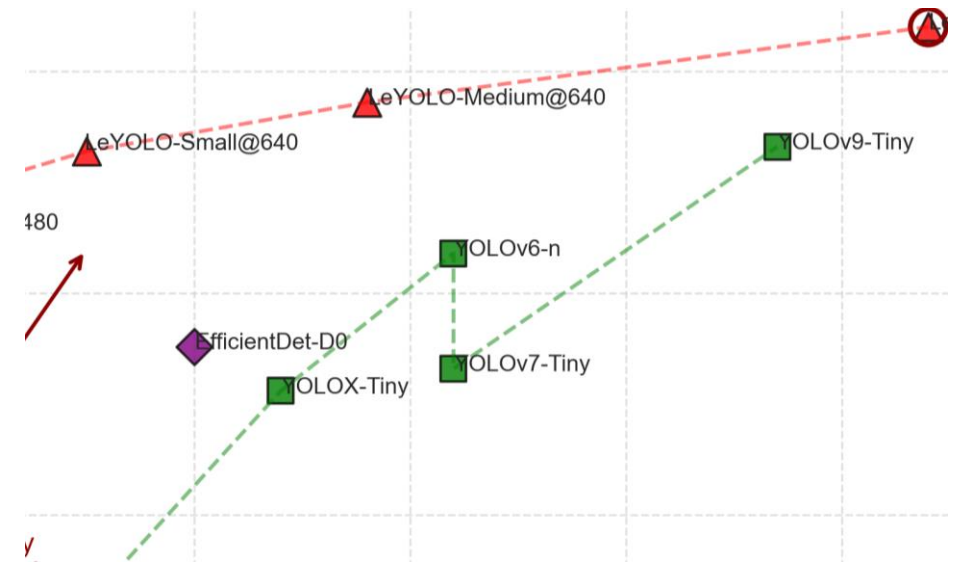
LeYOLO Small : 877 ms

LeYOLO Medium : **1039** ms

YOLOv8 Nano : **1235** ms

YOLOv9 Tiny : **1371** ms

YOLOv10 Nano : **1099** ms



Conclusion

- Réduire le coût de calcul des détecteurs d'objets
 - Battre l'état de l'art sur des jeux de données académiques
 - **LeYOLO** : Meilleur rapport paramètre / coût de calcul théorique que l'état de l'art des YOLO
 - **LeYOLO** : Meilleure vitesse d'exécution sur microcontrôleurs pour précision équivalente
 - **LeNeck** : Meilleur rapport paramètre / précision que SSDLite
- Intérêt applicatif fort : utilisation de nos recherches sur appareils embarqués pour nos partenaires industriels

Merci pour votre attention !

- A vos questions !



PS : Je cherche activement des rapporteurs et examinateurs pour ma thèse =)

