Improving YOLOv8 for Fast Few-Shot Object Detection by DINOv2 Distillation

International Conference on Image Processing **(ICIP)** 2025

Guillaume Fourret[1,2], Marc Chaumont[1,3], Christophe Fiorio[1], Gérard Subsol[1]

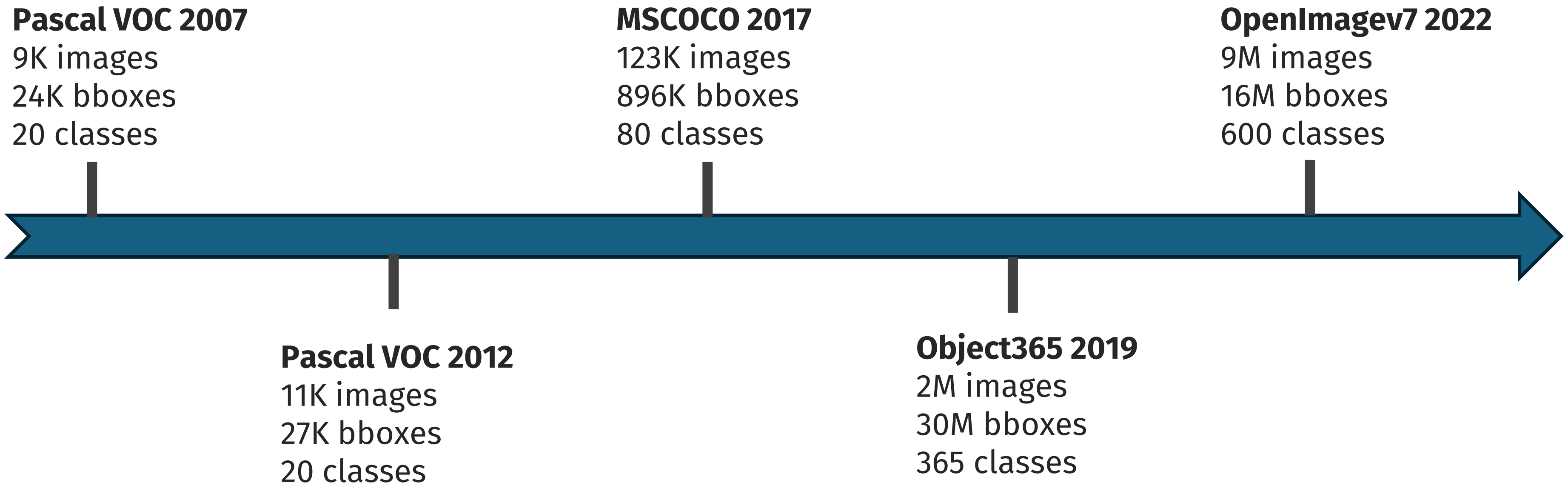
[1]ICAR, LIRMM, University of Montpellier, CNRS, France
[2]Drone Geofencing, Nîmes, France
[3]University of Nîmes, France

Contact: guillaume.fourret@lirmm.fr

1. Object Detection

Deep learning has seen great progress, notably in object detection, driven by scaling up:

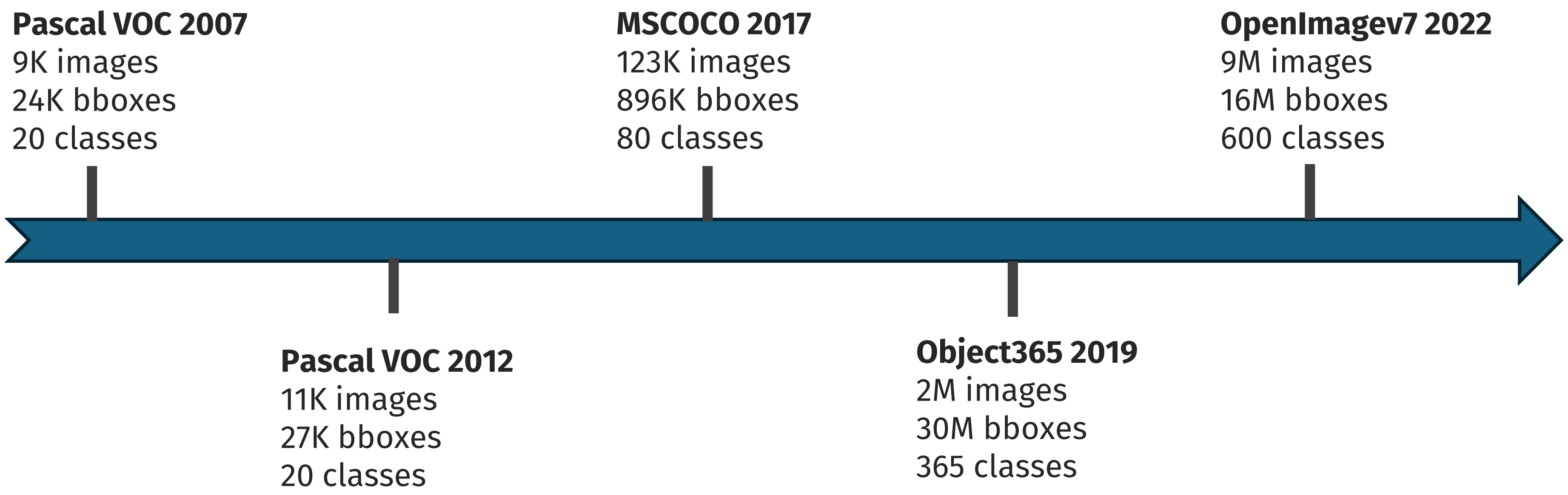
1. Model capacity (number of parameters)
2. Computational resources (number of GPUs)
3. Dataset size



1. Object Detection

Deep learning has seen great progress, notably in object detection, driven by scaling up:

1. Model capacity (number of parameters)
2. Computational resources (number of GPUs)
3. Dataset size



$\Rightarrow$ **What if we need real-time detection with few data (1–30 annotated boxes)?**

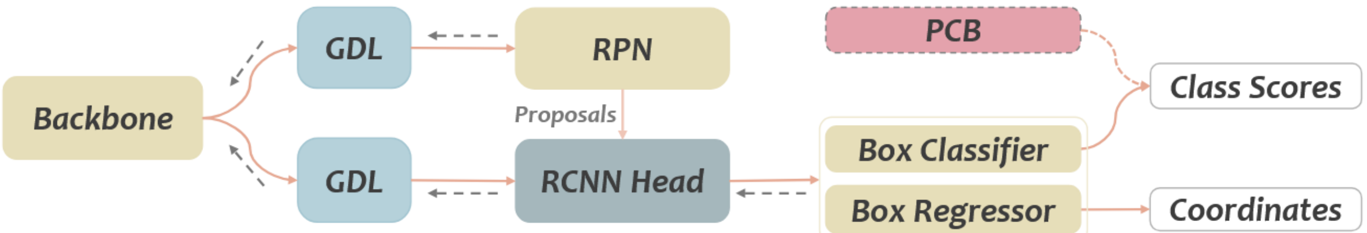
2. Few-Shot Object Detection (FSOD)

Goal of FSOD: Add to a detector new classes from only *K*-shot

Transfer learning / Meta learning
YOLO and Faster R-CNN

10-shot mAP on FSOD-adapted MSCOCO (source: paperswithcode)

Illustration of DeFRCN[1]



Decoupled Faster R-CNN Framework for Few-Shot Object Detection

[1]DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection, Limeng Qiao, et al, ICCV 2021
[2]Integrally migrating pre-trained transformer encoder-decoders for visual object detection, Feng Liu, et al, ICCV 2023
[3]DETReg: Unsupervised Pretraining with Region Priors for Object Detection, Amir Bar, et al, CVPR 2022

2. Few-Shot Object Detection (FSOD)

Goal of FSOD: Add to a detector new classes from only *K*-shot

Transfer learning / Meta learning
YOLO and Faster R-CNN

Transfer learning
ViT (+ Faster R-CNN) pretrained with SSL

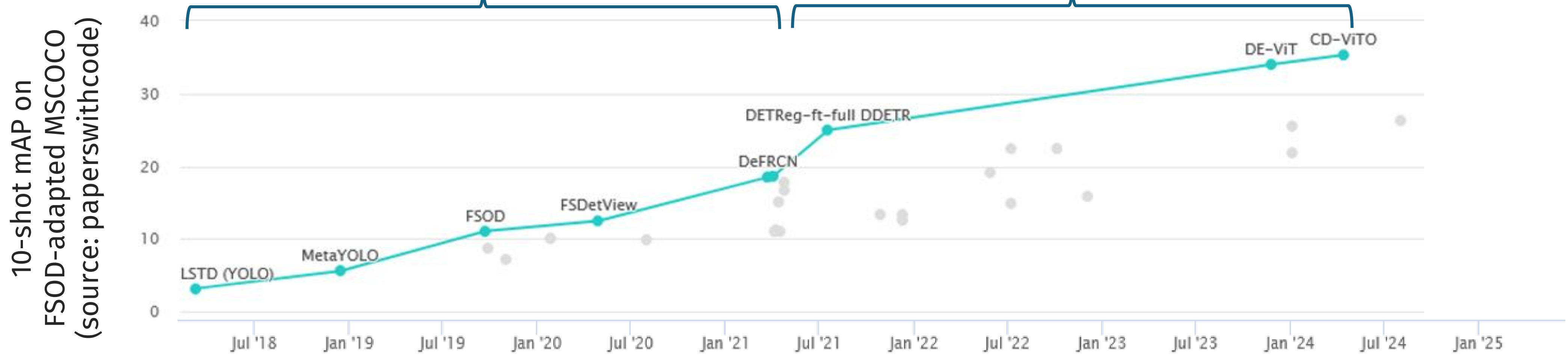
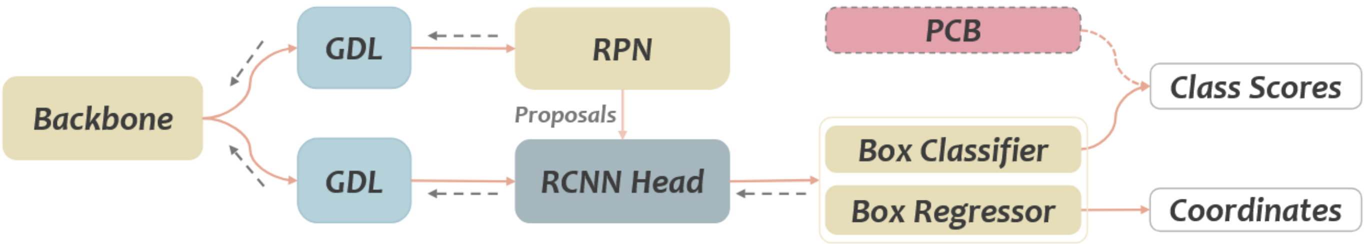


Illustration of DeFRCN[1]



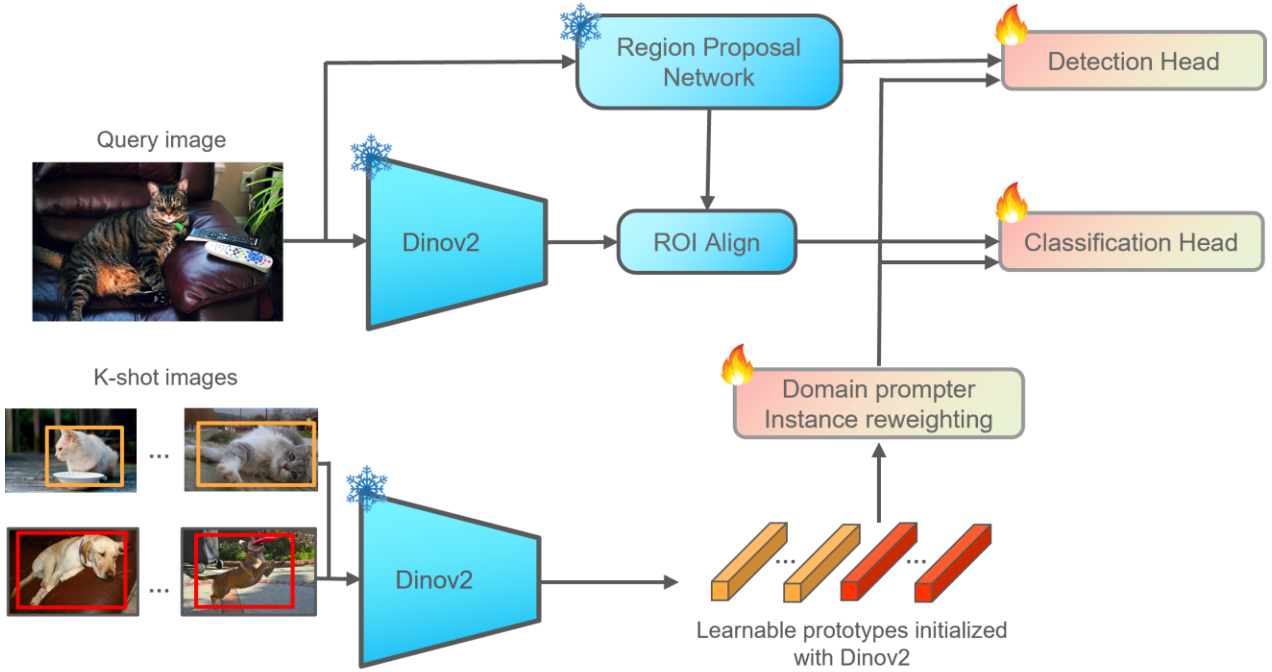
Decoupled Faster R-CNN Framework for Few-Shot Object Detection

⇒ Recently, **S**elf-**S**upervised **L**earning (**SSL**) based methods have shown better results[2,3]

[1]DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection, Limeng Qiao, et al, ICCV 2021
[2]Integrally migrating pre-trained transformer encoder-decoders for visual object detection, Feng Liu, et al, ICCV 2023
[3]DETReg: Unsupervised Pretraining with Region Priors for Object Detection, Amir Bar, et al, CVPR 2022

2. FSOD and SSL

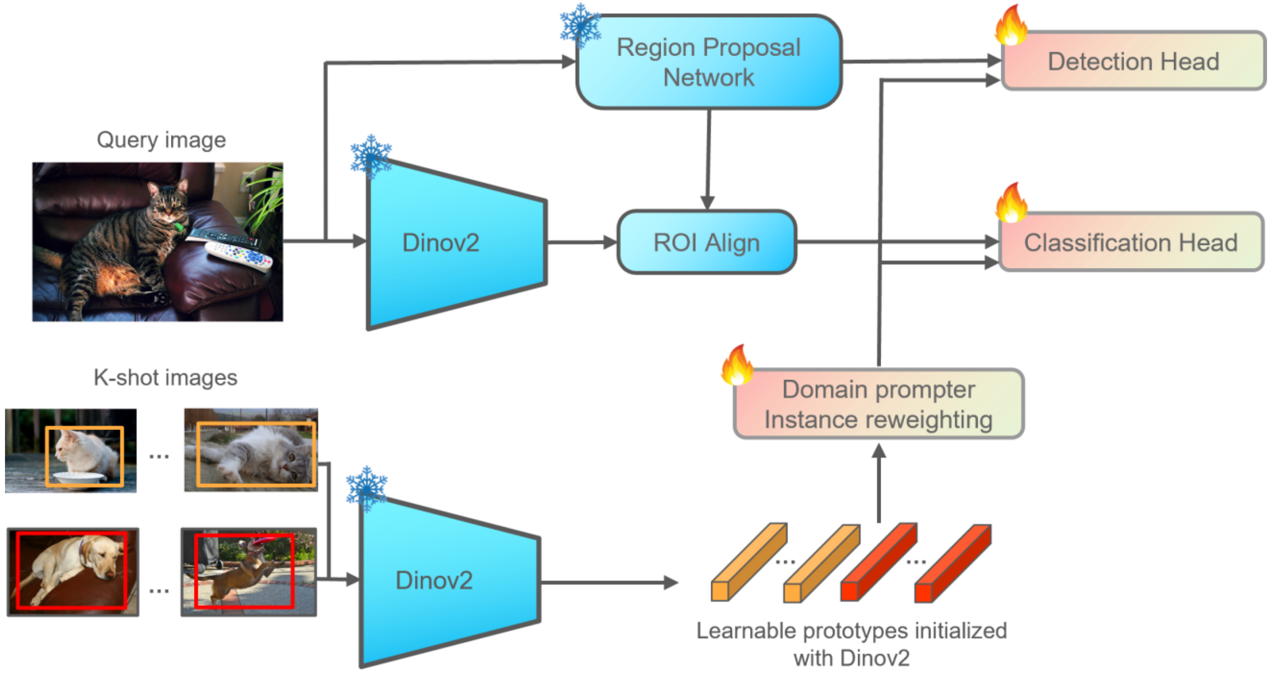
Recent FSOD methods as **FM-FSOD**[5] ,**DE-ViT**[6], and **CD-ViTO**[7] leverage foundation models like **DINOv2**[8,9]:



[5]Few-Shot Object Detection with Foundation Models, Guangxing Han, et al, CVPR 2024
[6]Detect Everything with Few Examples, Xinyu Zhang, et al, CoRL 2024
[7]Cross-Domain Few-Shot Object Detection via Enhanced Open-Set Object Detector, Yuqian Fu, et al, ECCV 2024
[8]DINOv2: Learning Robust Visual Features without Supervision, Maxime Oquab, et al, avril 2023
[9]Vision Transformers Need Registers, Timothée Darcet, et al, ICLR 2024

2. FSOD and SSL

Recent FSOD methods as **FM-FSOD**[5] ,**DE-ViT**[6], and **CD-ViTO**[7] leverage foundation models like **DINOv2**[8,9]:



- BUT far from real-time performance
- Meanwhile, YOLO series is the go-to for fast detection, yet not designed for FSOD

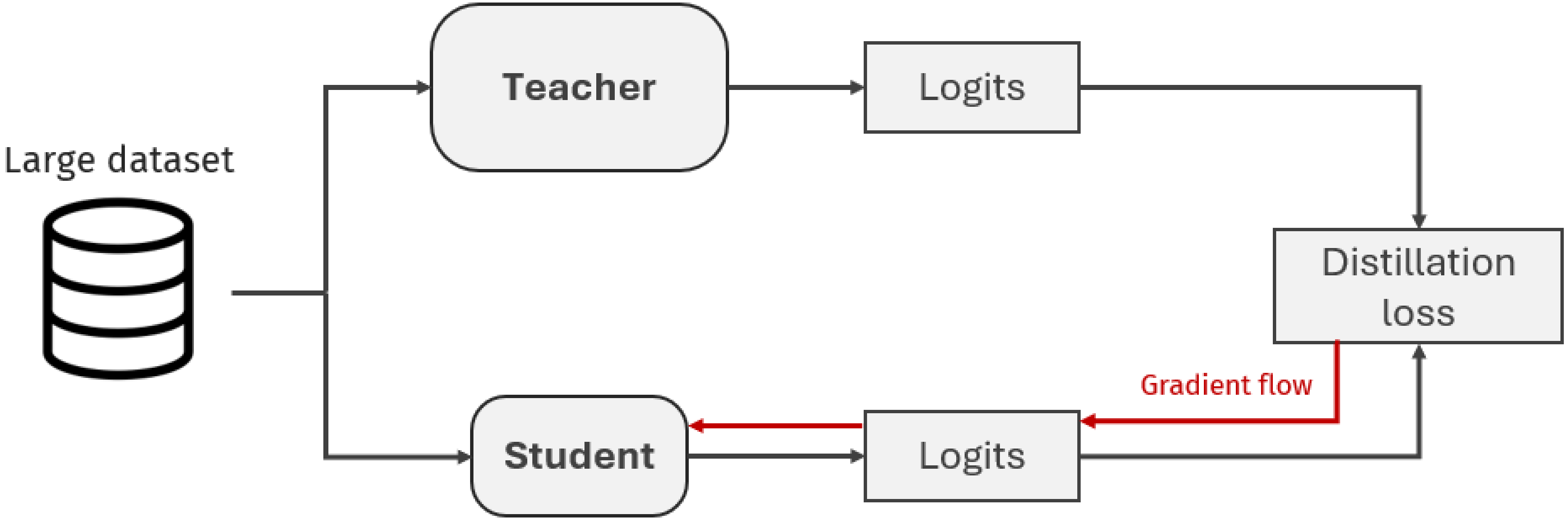
⇒ How to bring the capacity of **DINOv2** into **YOLO** for **real-time FSOD**?

[5]Few-Shot Object Detection with Foundation Models, Guangxing Han, et al, CVPR 2024
[6]Detect Everything with Few Examples, Xinyu Zhang, et al, CoRL 2024
[7]Cross-Domain Few-Shot Object Detection via Enhanced Open-Set Object Detector, Yuqian Fu, et al, ECCV 2024
[8]DINOv2: Learning Robust Visual Features without Supervision, Maxime Oquab, et al, avril 2023
[9]Vision Transformers Need Registers, Timothée Darcet, et al, ICLR 2024

3. A method of distillation for FSOD

Distillation[10] principle to obtain more efficient models :

1. Train a large teacher model on a huge dataset
2. Train a smaller student model (with same architecture) to mimic the teacher's predictions

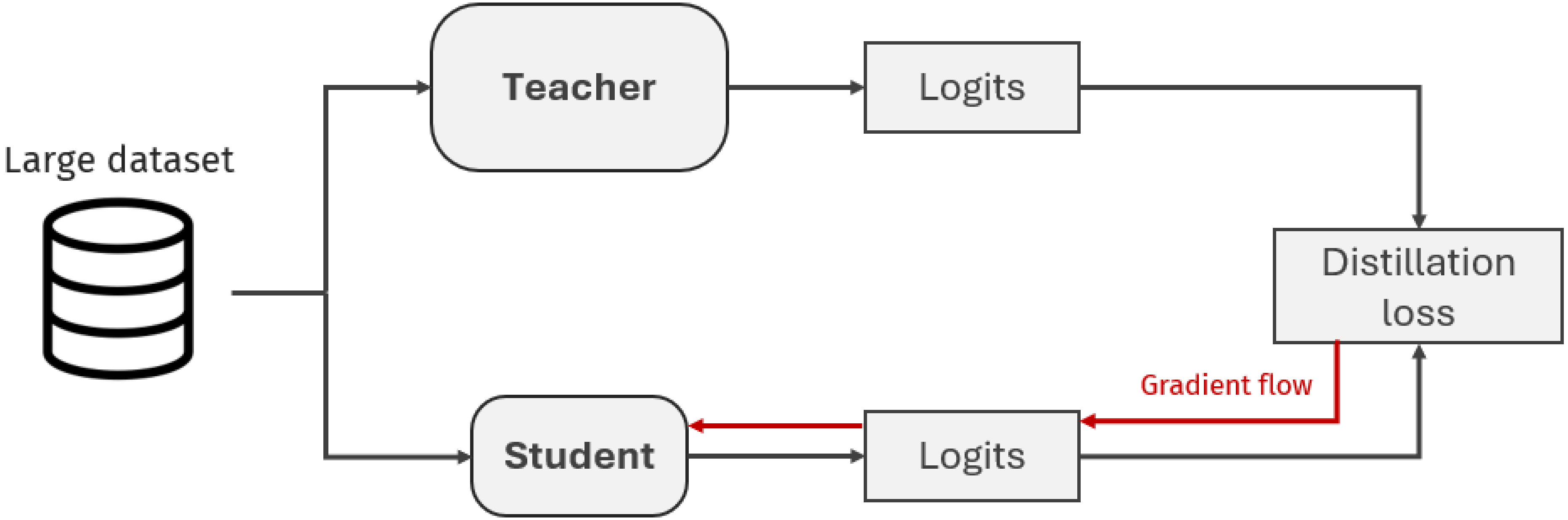


[10]Distilling the Knowledge in a Neural Network, Geoffrey Hinton, et al, 2015

3. A method of distillation for FSOD

Distillation[10] principle to obtain more efficient models:

1. Train a large teacher model on a huge dataset
2. Train a smaller student model (with same architecture) to mimic the teacher's predictions

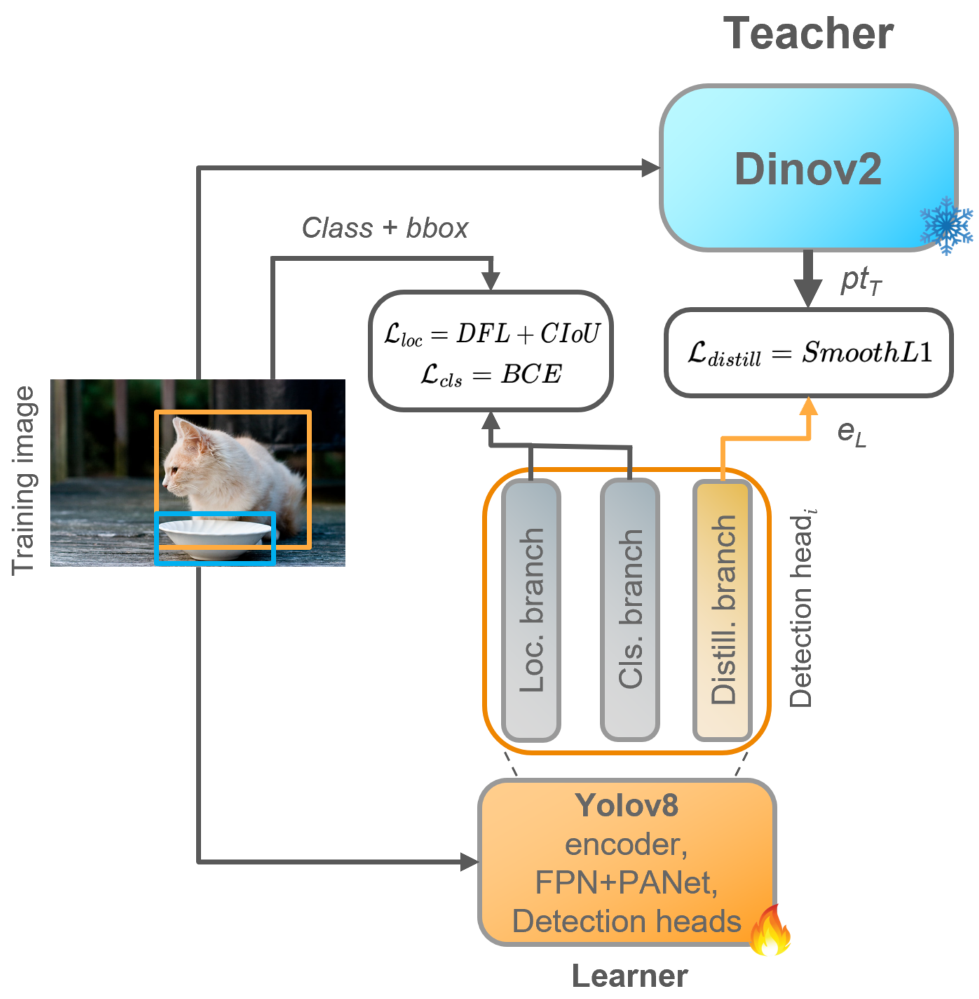


⇒ **How to perform distillation between YOLO and DINOv2?**

⇒ **Is distillation still relevant in FSOD setting?**

[10] Distilling the Knowledge in a Neural Network, Geoffrey Hinton, et al, 2015

3. Our distillation scheme: YOLOv8m_d1



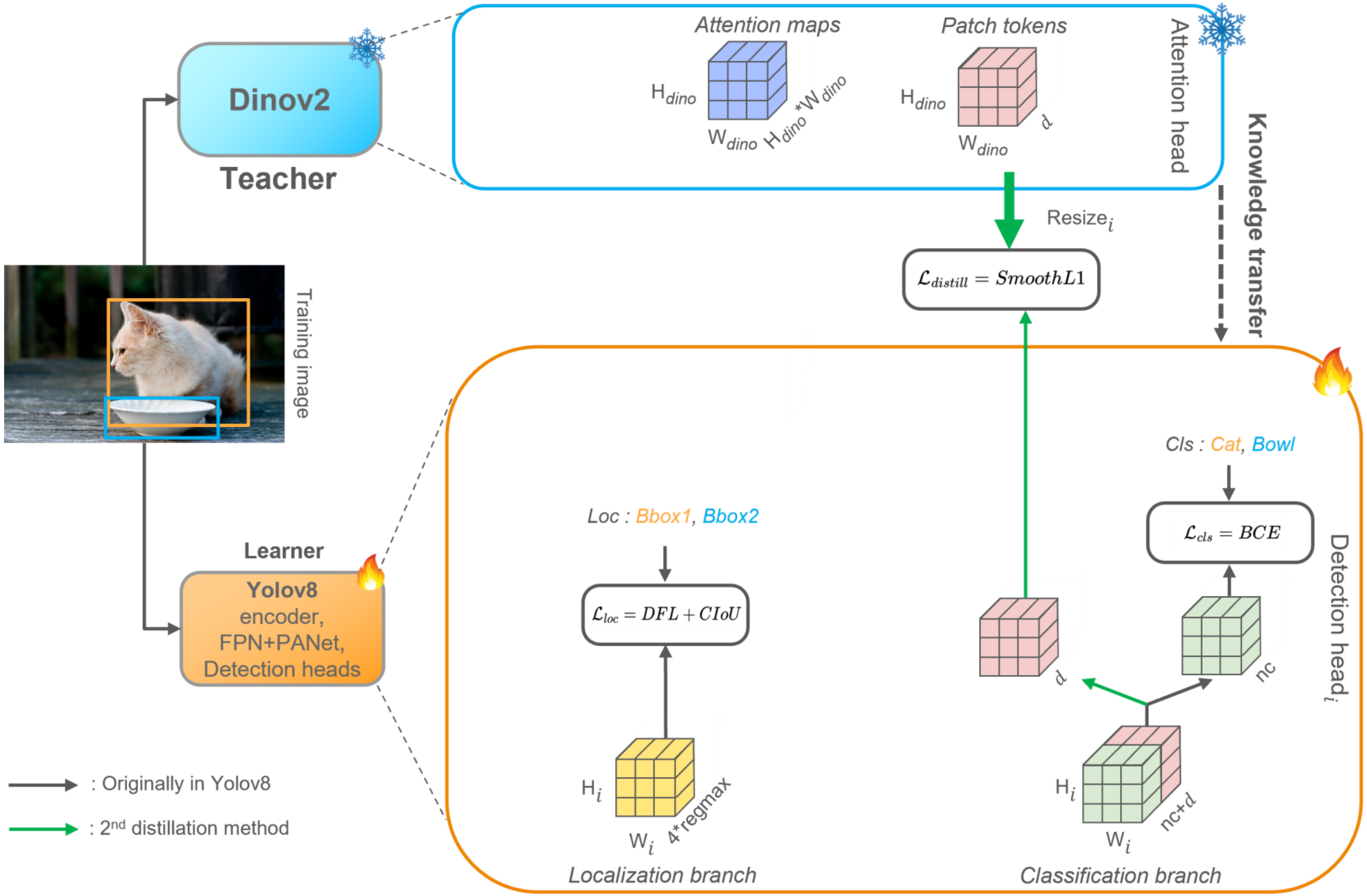
- YOLOv8 architecture employs parallel branches for classification and localization.
- **YOLOv8m_d1** introduces a new distillation branch.
- **DINOv2** serves as a frozen teacher.
- Distillation is performed by minimizing SmoothL1:

$$\mathcal{L}_{\text{SmoothL1}}(e_L, pt_T) = \begin{cases} 0.5 \cdot (e_L - pt_T)^2, & \text{if } |e_L - pt_T| < 1, \\ |e_L - pt_T| - 0.5, & \text{otherwise.} \end{cases}$$

Constraint: All additional parameters must be removable at inference time to ensure no impact on latency.

3. Our distillation scheme: YOLOv8m_d2

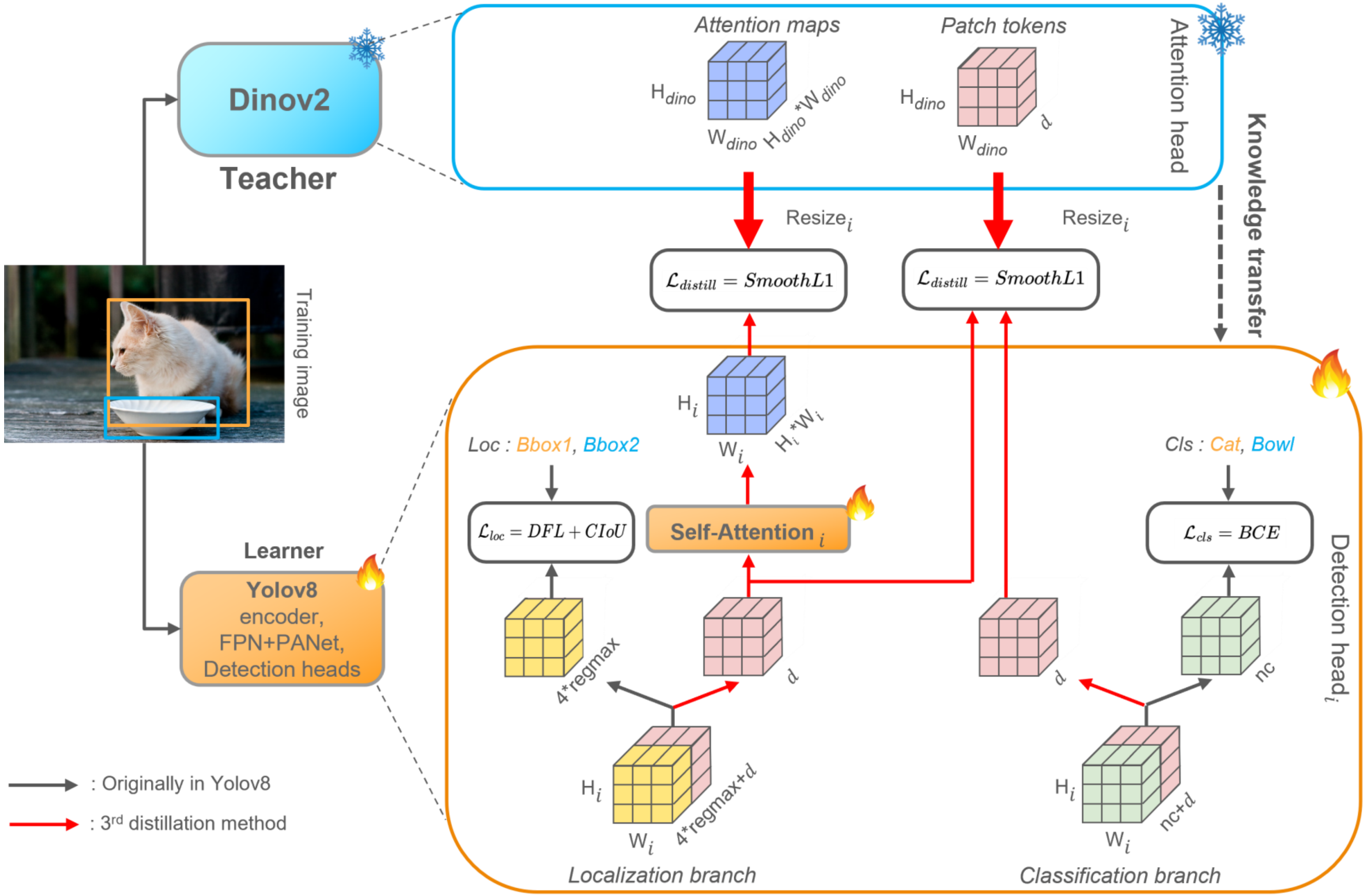
Broader distillation impact ⇒ Apply to classification branch



- **YOLOv8m_d2** removes this additional distillation branch
- Applies the distillation signal within the classification branch by extending last convolution

3. Our distillation scheme: YOLOv8m_d3

Broader distillation impact ⇒ Apply to both classification and localization branch



- **YOLOv8m_d3** keeps the distillation in the classification branch
- Also incorporates distillation in its localization with attention maps from DINOv2

4. Results on benchmarks

Benchmarked on the MSCOCO adapted to FSOD[11]:

60 classes for pretraining the models

Pre-training metrics	bAP50	bAP50:95
YOLOv8m vanilla	61.25	45.62
YOLOv8m_d1	62.40	46.61
YOLOv8m_d2	62.47	46.70
YOLOv8m_d3	62.44	46.15

[11] Frustratingly Simple Few-Shot Object Detection, Xin Wang, et al, ICML 2020
https://github.com/ucbdrive/few-shot-object-detection

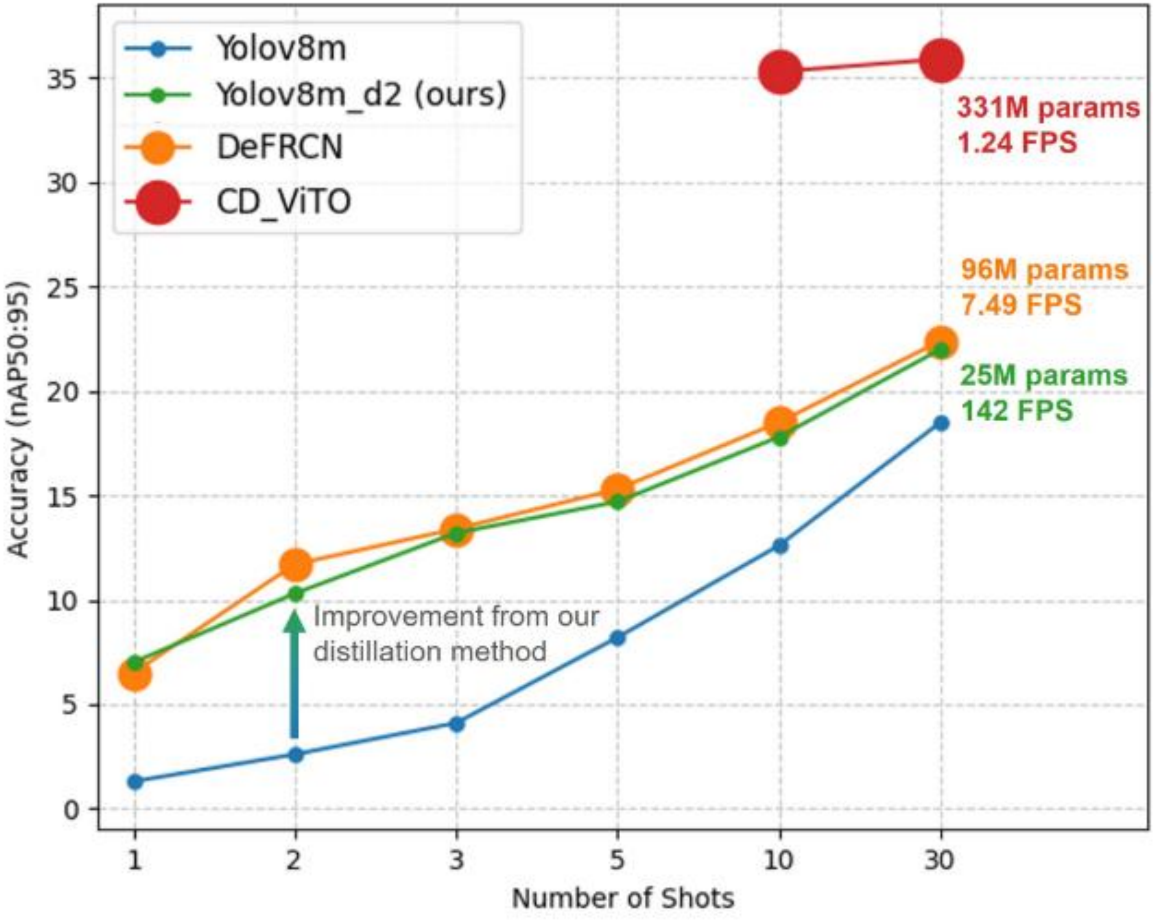
4. Results on benchmarks

Benchmarked on the MSCOCO adapted to FSOD[11]:

60 classes for pretraining the models

Pre-training metrics	bAP50	bAP50:95
YOLOv8m vanilla	61.25	45.62
YOLOv8m_d1	62.40	46.61
YOLOv8m_d2	62.47	46.70
YOLOv8m_d3	62.44	46.15

20 classes for the *K*-shot finetuning



[11]Frustratingly Simple Few-Shot Object Detection, Xin Wang, et al, ICML 2020
https://github.com/ucbdrive/few-shot-object-detection

4. Results on benchmarks

Benchmarked on the MSCOCO adapted to FSOD[11]:

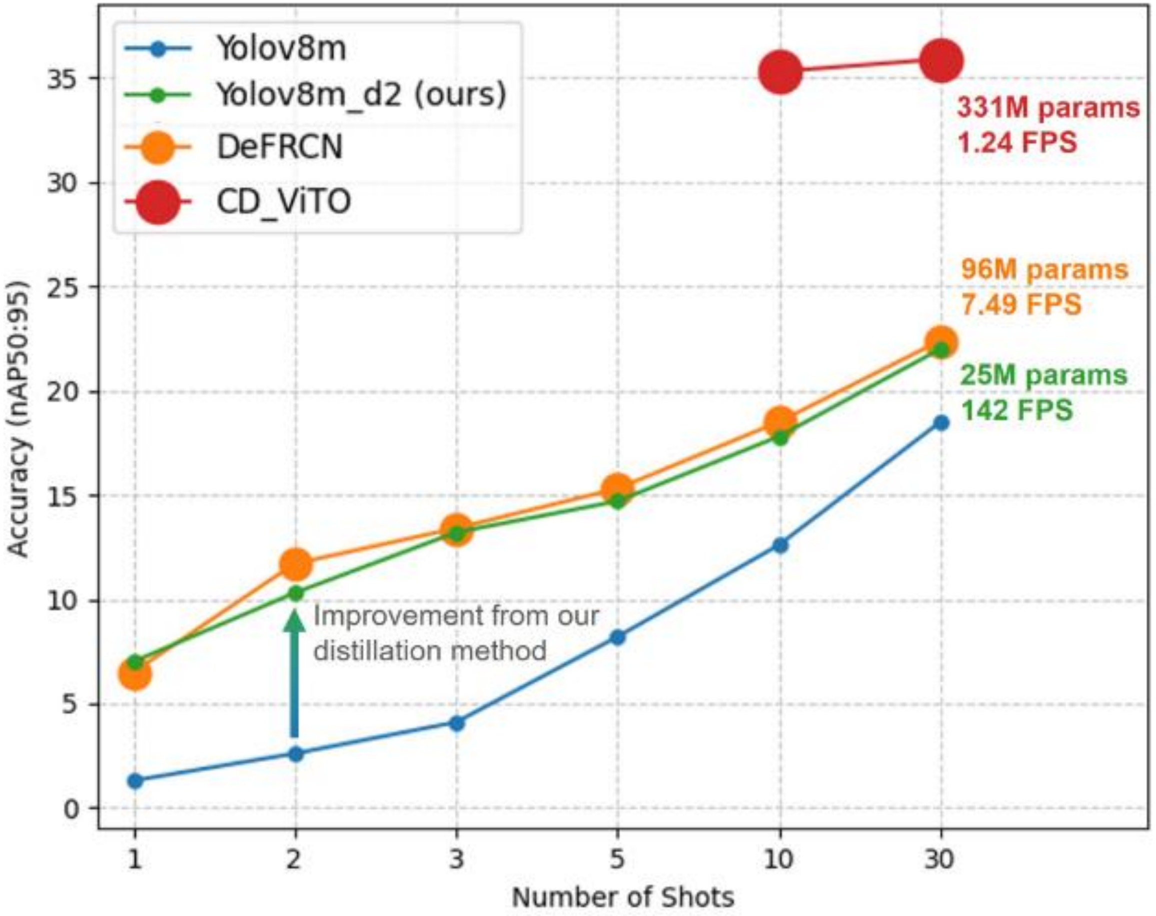
60 classes for pretraining the models

Pre-training metrics	bAP50	bAP50:95
YOLOv8m vanilla	61.25	45.62
YOLOv8m_d1	62.40	46.61
YOLOv8m_d2	62.47	46.70
YOLOv8m_d3	62.44	46.15

While being a lot more lightweight and faster !

	Number of Parameters ↓	FPS ↑
YOLOv8m	25,902,640	142
DeFRCN	96,754,958	7.49
CD-ViTO	331,149,640	1.24

20 classes for the *K*-shot finetuning



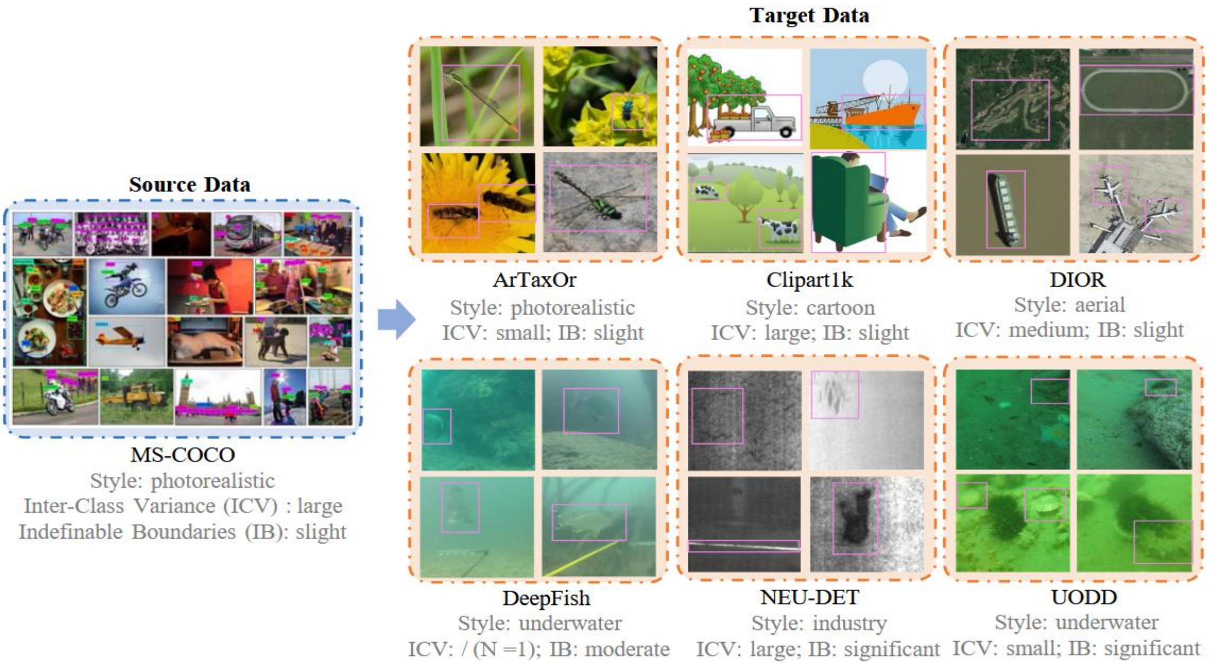
⇒ **Yolov8m_d1** performs a bit worse than others
⇒ **Yolov8m_d2** best from 1 to 5 shots
⇒ **Yolov8m_d3** best from 10 to 30 shots

[11]Frustratingly Simple Few-Shot Object Detection, Xin Wang, et al, ICML 2020
https://github.com/ucbdrive/few-shot-object-detection

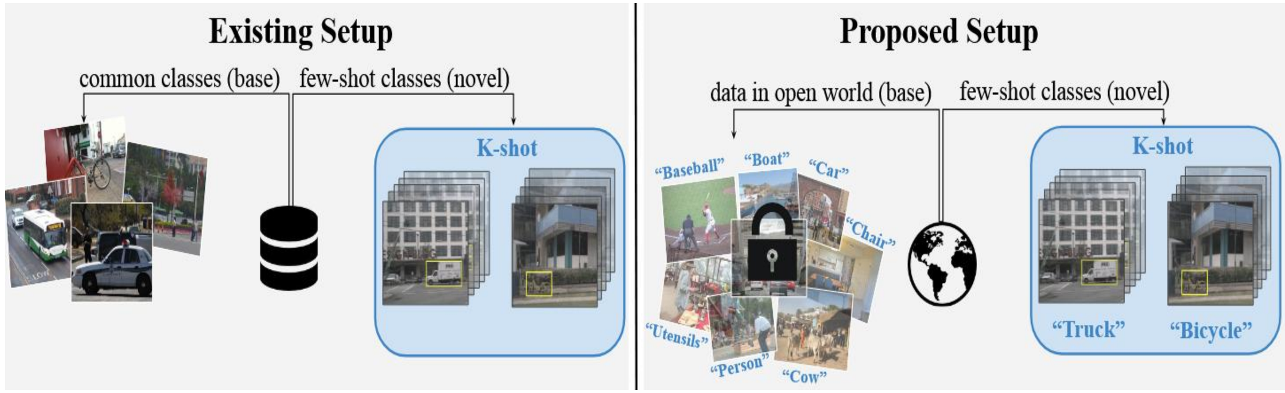
5. Conclusion

- We introduced distillation-based supervision to benefits from SSL-pretrained extractor in a fast detector
- Further work could investigate results on Cross-Domain FSOD[12] and comparisons to VLM and zero-shot methods[13,14]

Discrepancy between domain in CD-FSOD



Setup for VLM evaluation in FSOD



[12] Cross-domain few-shot object detection via enhanced open-set object detector, Yuqian Fu, et al, ECCV 2025
[13] Revisiting Few-Shot Object Detection with Vision-Language Models, Anish Madan, et al, NeurIPS2024
[14] Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, Shilong Liu, et al, ECCV2024