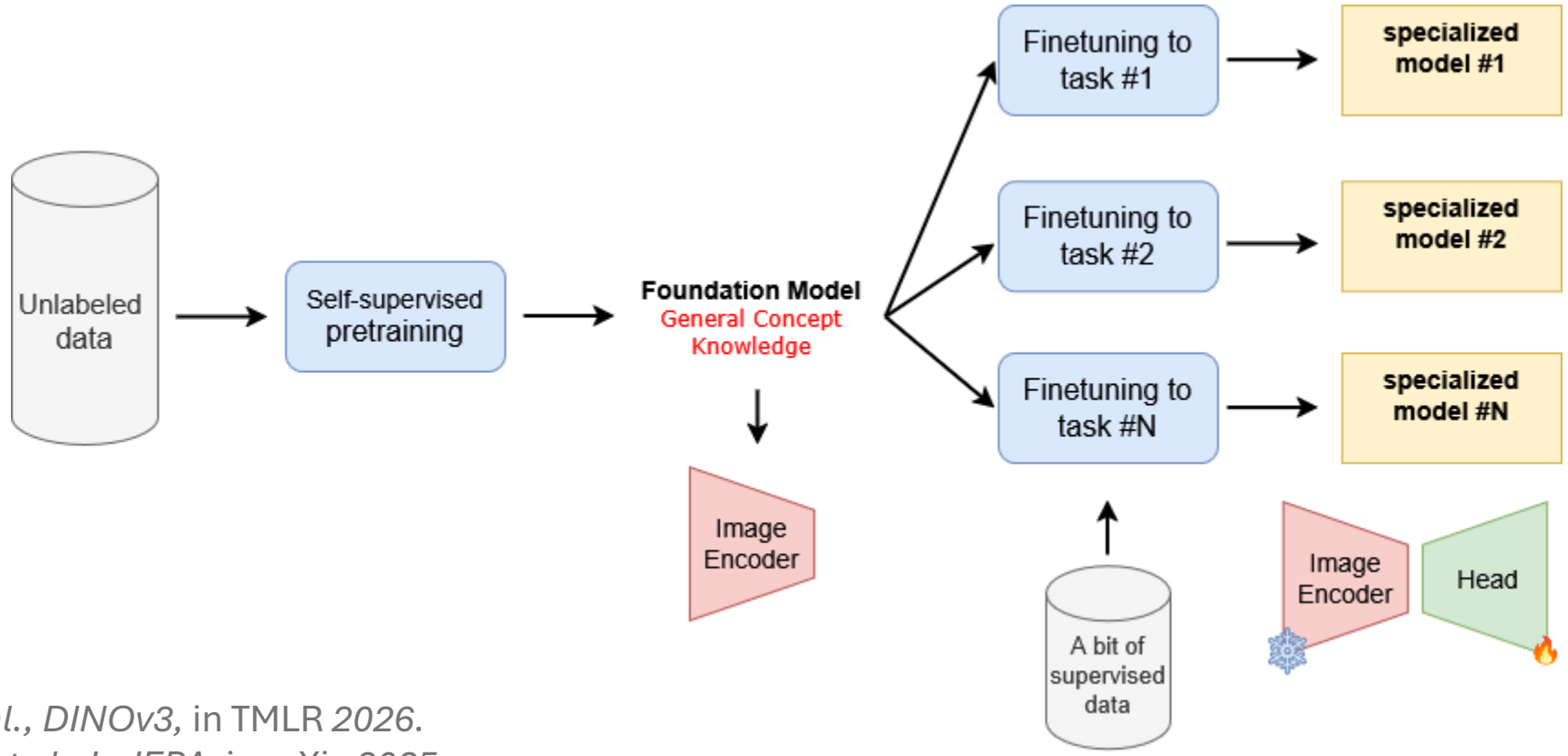
SAM2 R-CNN: Transferring SAM 2 Knowledge For Data Efficient Instance Segmentation

Mehdi Gharbage [1,2], Céline Teulière [1], Pierre Bouges [2] & Thierry Chateau [1]

[1] Institut Pascal, CNRS, Clermont-Ferrand, France

[2] Michelin

The Era of Foundation Models: General concept representations for any vision task



Simeoni *et al.*, *DINOv3,* in TMLR *2026*.
Balestriero *et al.*, *LeJEPA,* in arXiv *2025*.

Web-Pretrained Vision Foundation Models Do Not Survive Semantic Shift

- Representations learned on natural images transfer poorly to out-of-distribution (OOD) data (*e.g.*, industrial images) [1]
- Large-scale self-supervised VFM (*e.g.*, DINOv3) rely more on object shape than object texture [2].

➢ Is there a VFM that transfers better to OOD Instance Segmentation ?

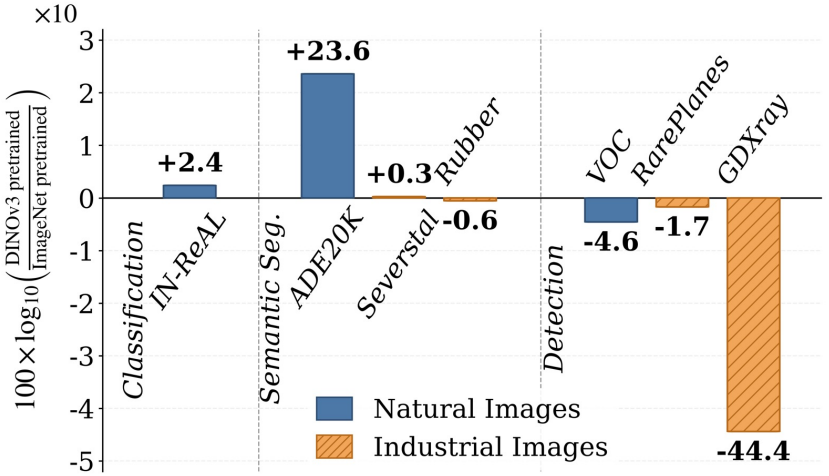


Fig. 1 : ***Natural vs. Industrial Images.*** Relative performance gap between DINOv3 and ImageNet supervised classification pretraining with ConvNets backbones. [1]

[1] Gharbage et *al., Rethinking Transfer Learning for Industrial Inspection, in* CVPR Workshop 2026.
[2] Park et *al., What Do Self-Supervised Vision Transformers Learn?,* in ICLR 2022

Emerging properties in SAM2 : A Transferable representation under semantic shift

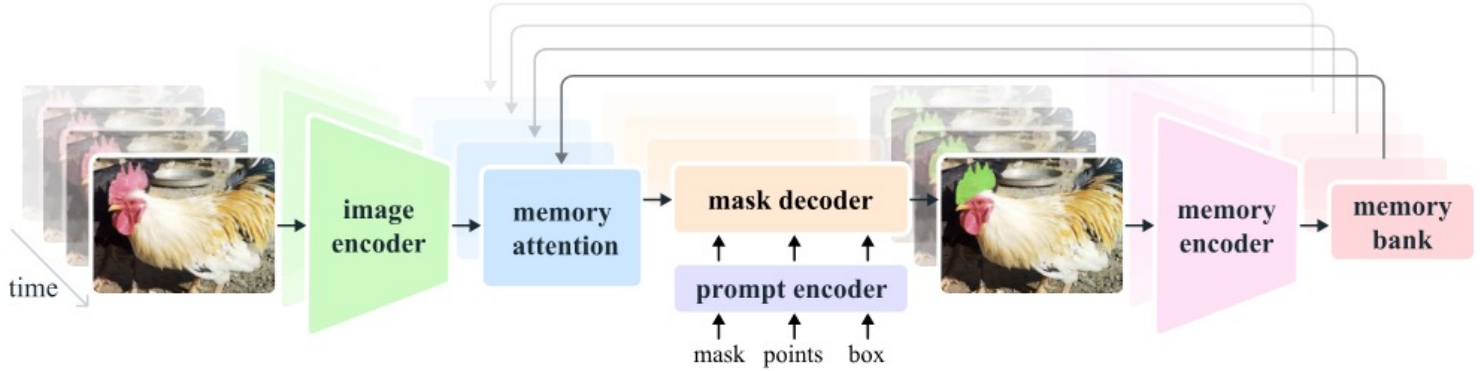


Fig. 1 : SAM 2 architecture. [2]

- Promptable Image & Video Object Segmentation & Tracking
- Supervised & Auto-Supervised class-agnostic mask prediction
- +50k videos & +35M masks

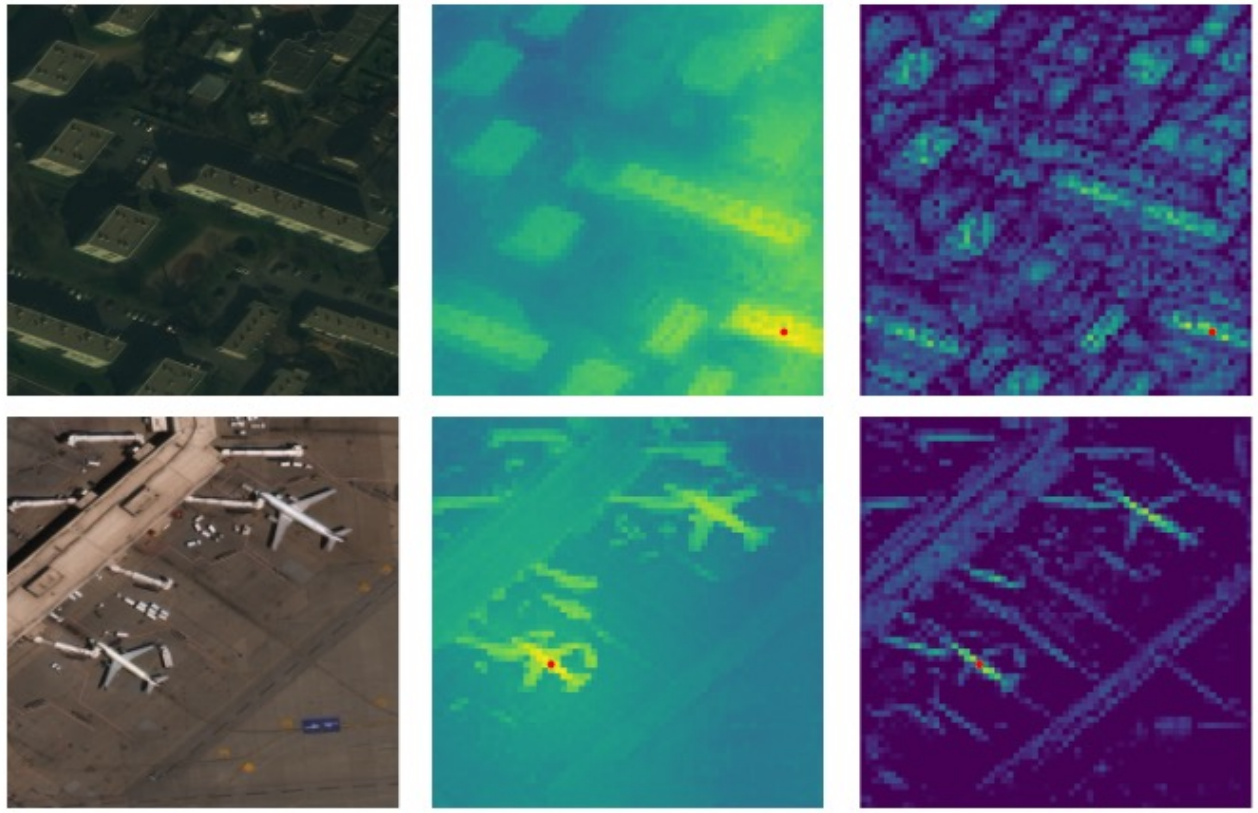
➢ **A visual similarity prior emerged** [2]

Fig. 2 : Latent space probing of SAM. [1]

[1] Kirillov et *al.*, *Segment Anything, in* ICCV 2023.
[2] Cuttano et *al.*, *SANSA: Unleashing the Hidden Semantics in SAM2 for Few-Shot Segmentation,* in NeurIPS 2025.

Emerging properties in SAM2 : A Transferable representation under semantic shift

➢**A visual similarity prior emerged**



SAM 2 Image Representation is Not Directly Transferable

- “SAM2 does encode semantic information, which is however entangled with instance-specific feature optimized for object tracking” [1].

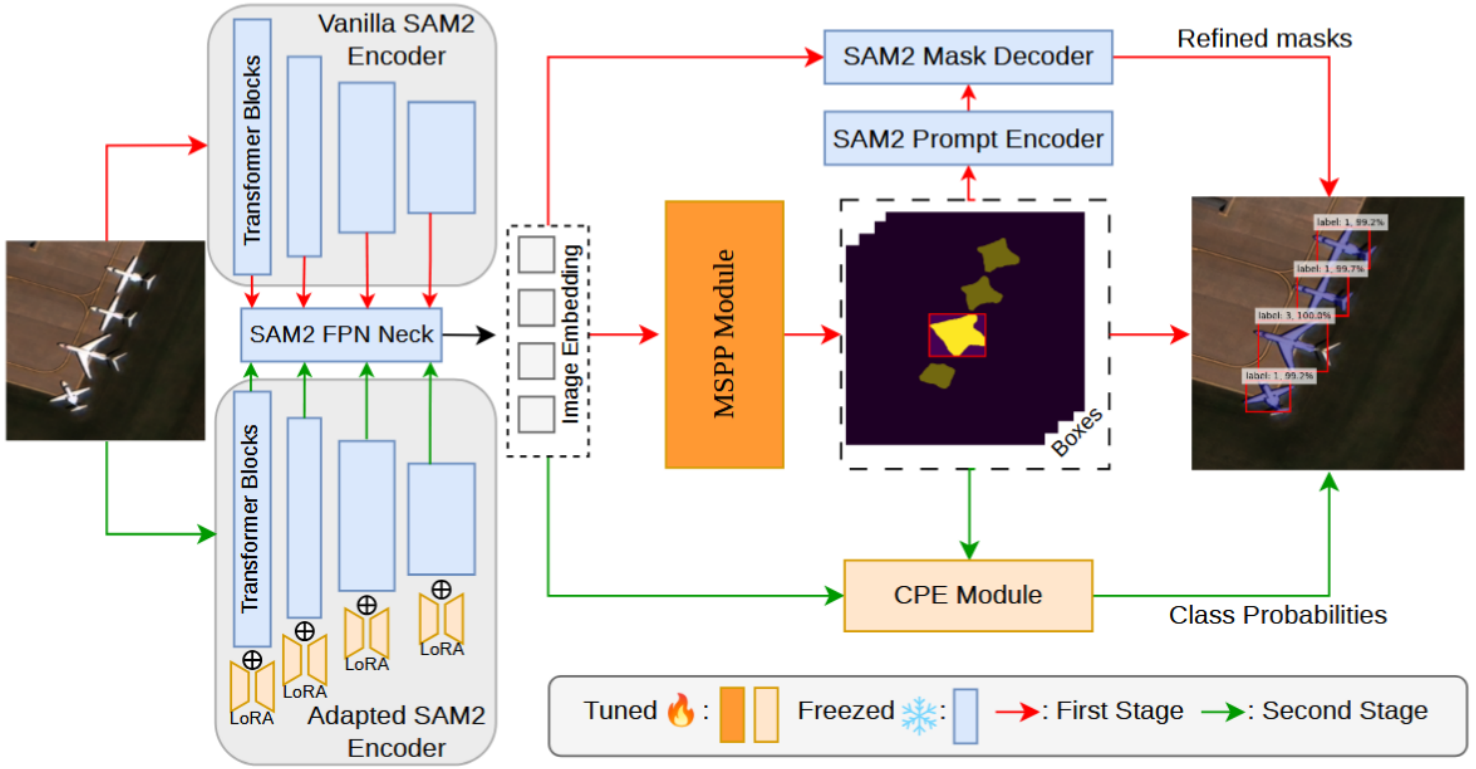
Model	Top-1 Acc. (%)	Top-5 Acc. (%)
SAM	11.06	25.37
SAM 2	23.16	44.44
CLIP	73.92	92.89
DINOv2	77.43	93.78

Fig. 1 : ImageNet classification accuracy under linear probing of foundation models as image encoders. [2]

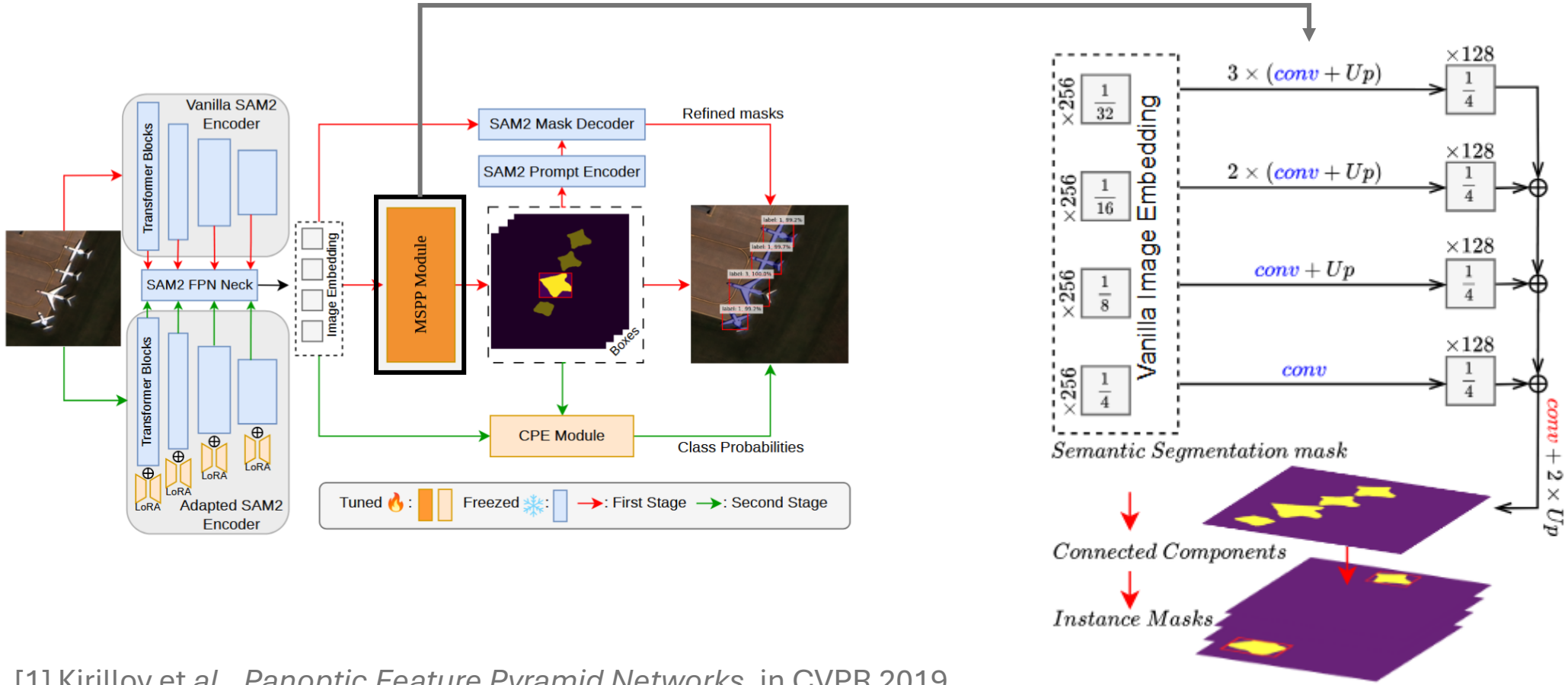
[1] Cuttano et *al., SANSA: Unleashing the Hidden Semantics in SAM2 for Few-Shot Segmentation,* in NeurIPS 2025.
[2] Espinosa et *al., There is no SAMantics! Exploring SAM as a Backbone for Visual Understanding Tasks,* in arXiv 2024.

SAM2 R-CNN: Unlocking SAM 2 Semantics for data-efficient OOD Instance Segmentation

- A two-stage framework:
  - ➢ Automatic Prompting
  - ➢ Embedding SAM 2 with semantic knowledge

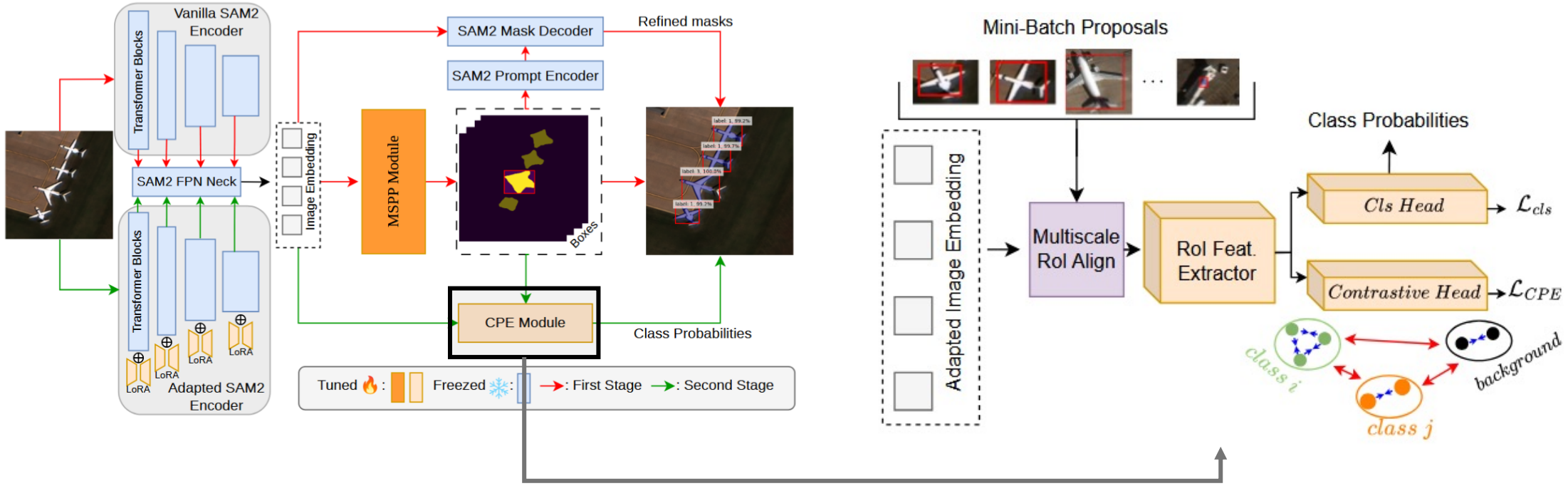


SAM2 R-CNN: Unlocking SAM 2 Semantics for data-efficient OOD Instance Segmentation



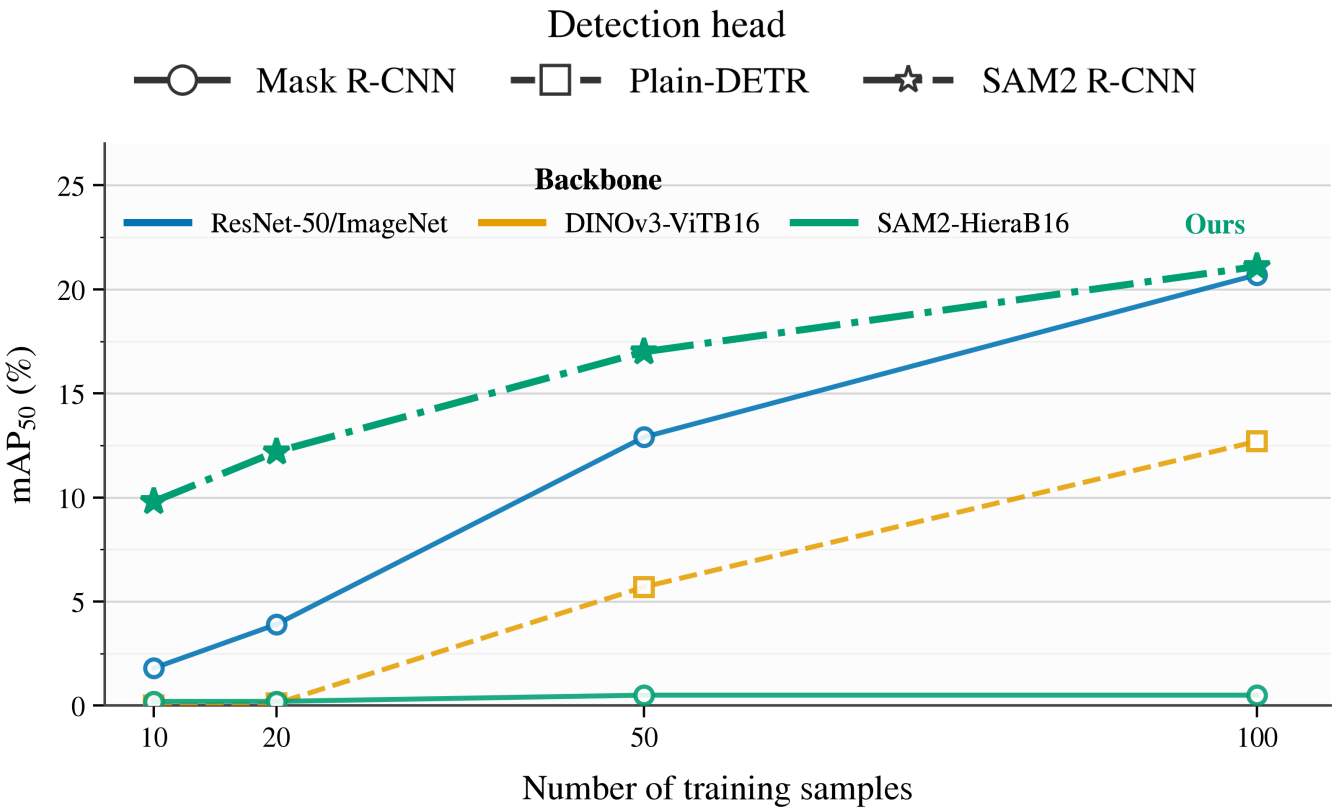
[1] Kirillov et *al.*, *Panoptic Feature Pyramid Networks,* in CVPR 2019.

SAM2 R-CNN: Unlocking SAM 2 Semantics for data-efficient OOD Instance Segmentation



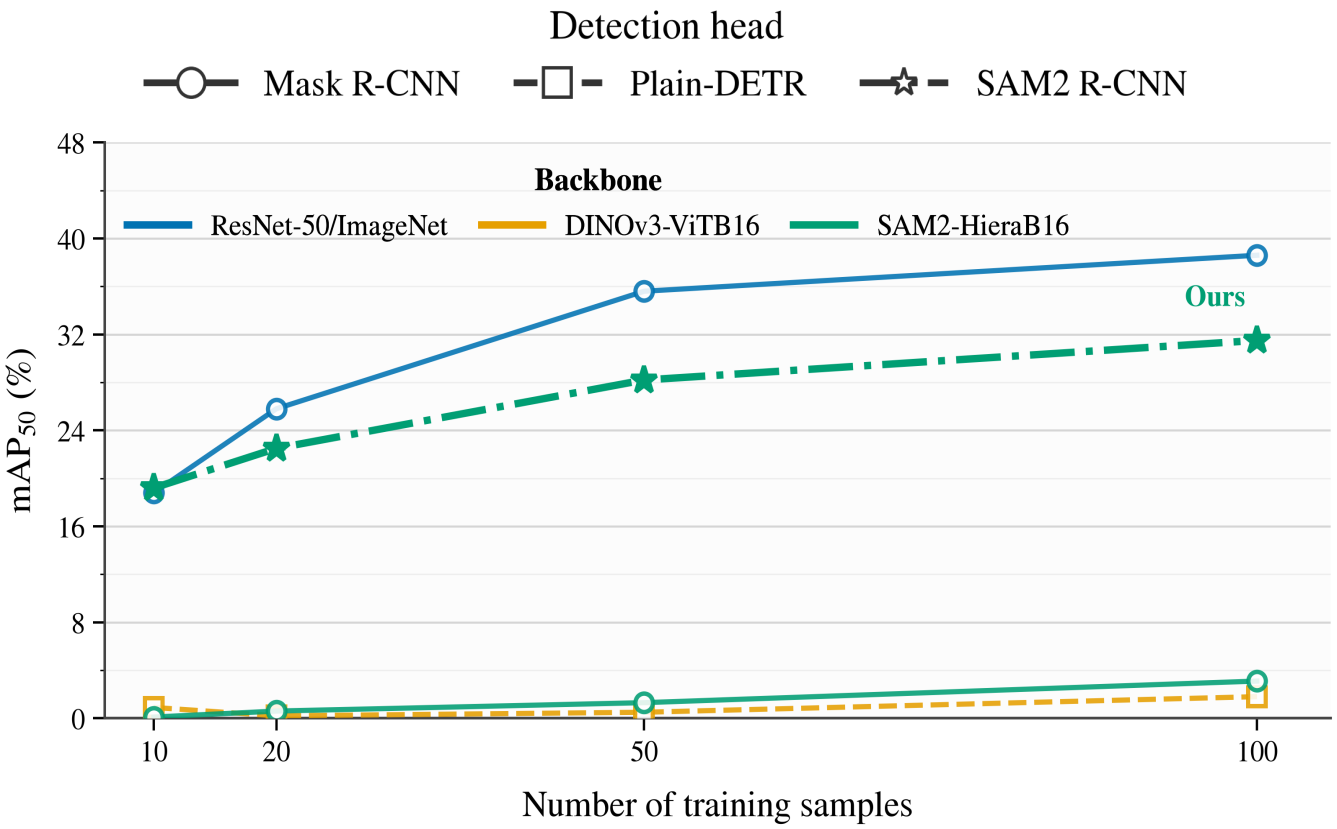
[1] Sun et *al.*, *FSCE: Few-shot object detection via contrastive proposal encoding,* in CVPR 2021.

Experimental results on RarePlanes dataset for Aircraft localization



- SAM 2 really lacks semantics, even with LoRA adaptation.
- DINOv3 with a Plain-DETR head needs some more data
- SAM2-RCNN is how we can leverage SAM 2 embeddings for OOD instance segmentation.

Experimental results on Spacenet-v2 for building localization: The failure case



- SAM 2 R-CNN **struggles in dense crowded scenes.**

SAM 2 R-CNN's performance gains come at a cost

- Equivalent number of trainable parameters compared to standard Mask R-CNN (Tab. 1).
- SAM2 R-CNN comes at a **high computational cost**, more than doubling both training and inference time.

➢ Is the second semantic stage necessary?

Method	Contrastive loss	LoRA finetuning	10
Mask R-CNN	-	✗	0.89 ± 0.78
	-	✓	1.21 ± 0.33
SAM2 R-CNN	✗	✗	11.80 ± 2.69
	✓	✗	11.36 ± 2.96
	✓	✓	12.26 ± 3.78

+ 10,6

+ 0,46

Tab. 3: Ablation study on Aircraft detection on 10 training samples (mAP).

Method	Comp. (M)			Total
	Prop.	Head	Adapt.	
Mask R-CNN	0.6	16.5	–	17.1
SAM2 R-CNN	0.9	15.1	2.0	18.0

Tab. 1: # Trainable parameters

Method	Train GPU-h	Infer. s/img ↓
Mask R-CNN (ResNet-50)	0.6	0.37
Mask R-CNN (Hiera-B)	2.7	0.85
SAM2 R-CNN	6.4	1.79

x2

Tab. 2: Runtime cost

Take-away and perspectives

- **SAM2 R-CNN** leverages **SAM 2's visual similarity prior** for few-shot instance segmentation **under semantic domain shift**.
- It shows **promising few-shot performance**, but **struggles in dense or crowded scenes** due to its bottom-up proposal strategy.
- Its computational cost limits its use as a fully end-to-end detector, making it more suitable as an annotation assistant or low-data bootstrapping tool.

Move beyond adapting generic models: build foundation models for the target domain!

Thank you

Give it a try