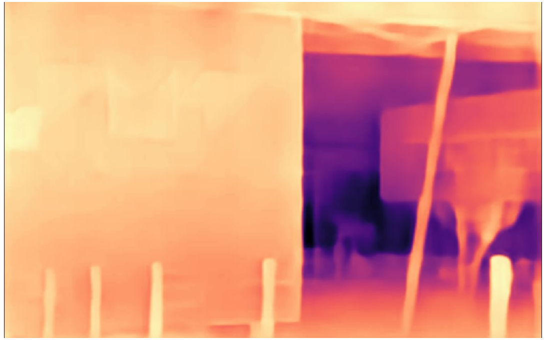
Foundation Models for Vision: DINOv2 & SAM

Équipe BRAIn

Content

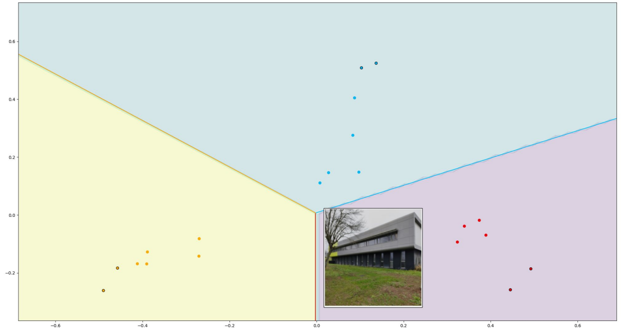


DINOv2

a model to produce universal features

SAM

a model to segment any scene

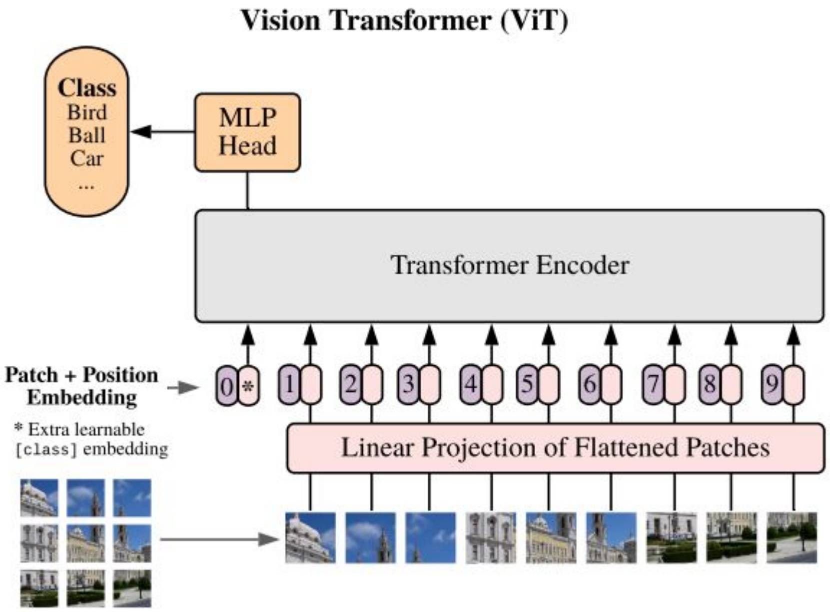


Hands on:

few-shot classification with DINOv2

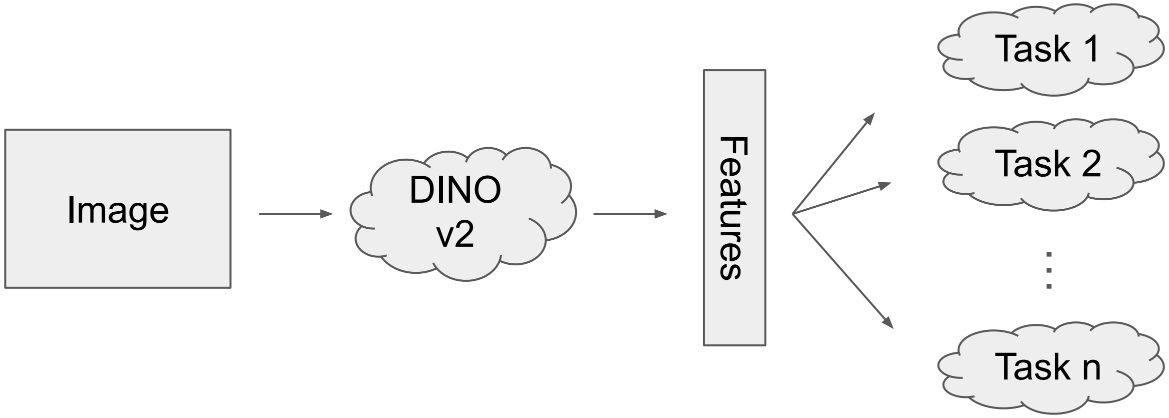
SoA architectures on existing “curated” datasets

- Vision transformers
- Tokenization of images
- Addition of a “Position Embedding”
- Addition of a CLS token: synthesis of patches



Purpose of a Foundation Model for Vision

From DINOv2 introduction: “producing all-purpose visual features, i.e., features that work across image distributions and tasks **without finetuning**”



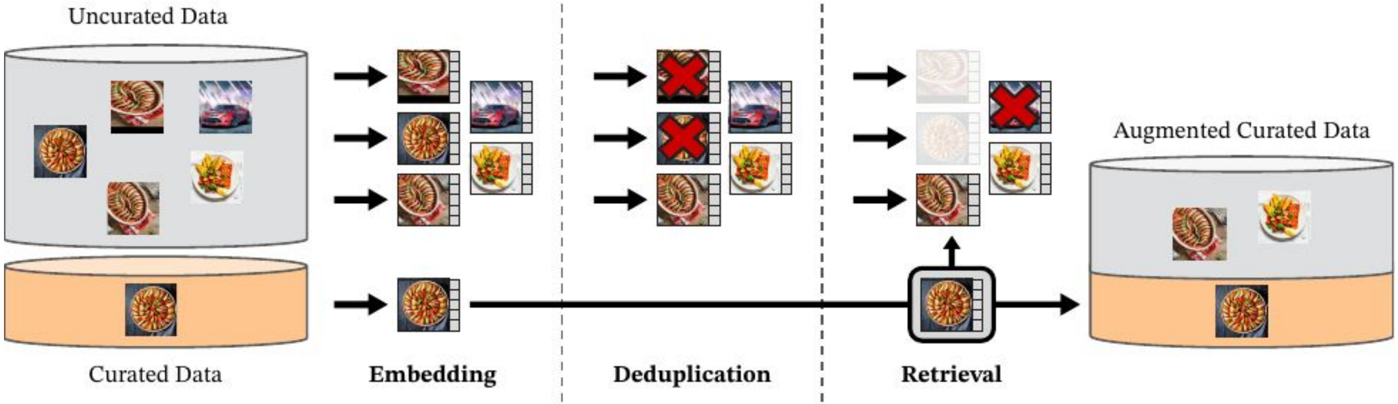
What are the challenges to build such a system ?

- Self-supervised training methods exist, but either on **small curated dataset (Imagenet-1K)** or on **big uncurated dataset**
  - Curated data : quality, diversity, balance
  - Adaptation to specific tasks are performed through fine tuning
  - In order to provide the best pretrained encoders, need to train with **big curated data**

- Revisit and combine methods to **scale** on **data** and **model size**
  - Stabilise training
  - Accelerate

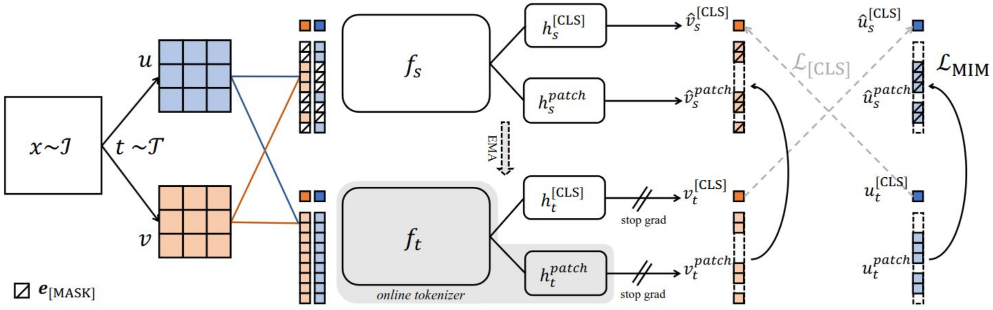
Create a “curated dataset”

- First, collection of “curated data” (ImageNet-1k & ImageNet-22k & others)
- A collection of unfiltered “uncurated data”: 1.2B images
- Goal: retrieve images that are close to curated datasets
- These are curated through Embedding / Deduplication / Retrieval
    - Deduplication from Pizzi et al.
    - Cosine similarity and clustering
- To create the LVD-142M dataset



Scaling data: how to train with so many images

- Self-supervision
  - Image-level objective (through EMA)
  - Patch-level objective
- Untying head weights for both aforementioned objectives
- Adapting the resolution: high resolution during training for downstream tasks



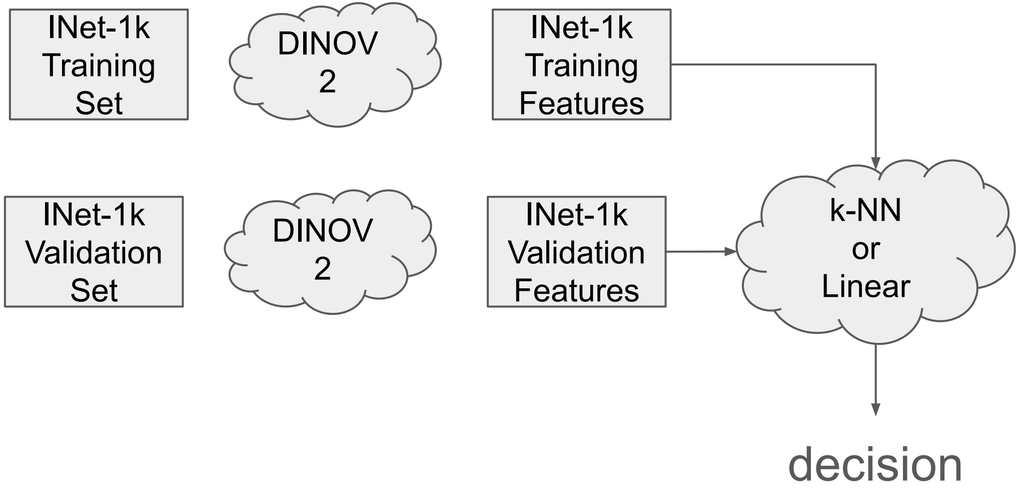
Scaling: training efficiency

- Flash Attention
  - Efficient Tiling
- Nested tensors
  - Different crop version in the same forward pass
- Efficient Stochastic Depth
  - Take full benefit from stochastic depth
- Fully Sharded Data Distribution
  - Spreading replicas across GPUs - teacher, students, optimizer moments
- Model Distillation
  - To target smaller models

model	# of params	with registers	ImageNet k-NN	ImageNet linear	download
ViT-S/14 distilled	21 M	❌	79.0%	81.1%	backbone only
ViT-S/14 distilled	21 M	✅	79.1%	80.9%	backbone only
ViT-B/14 distilled	86 M	❌	82.1%	84.5%	backbone only
ViT-B/14 distilled	86 M	✅	82.0%	84.6%	backbone only
ViT-L/14 distilled	300 M	❌	83.5%	86.3%	backbone only
ViT-L/14 distilled	300 M	✅	83.8%	86.7%	backbone only
ViT-g/14	1,100 M	❌	83.5%	86.5%	backbone only
ViT-g/14	1,100 M	✅	83.7%	87.1%	backbone only

Experiences & Results

- On ImageNet-1k classification
  - Use CLS token
  - Downstream classification with k-NN or Linear

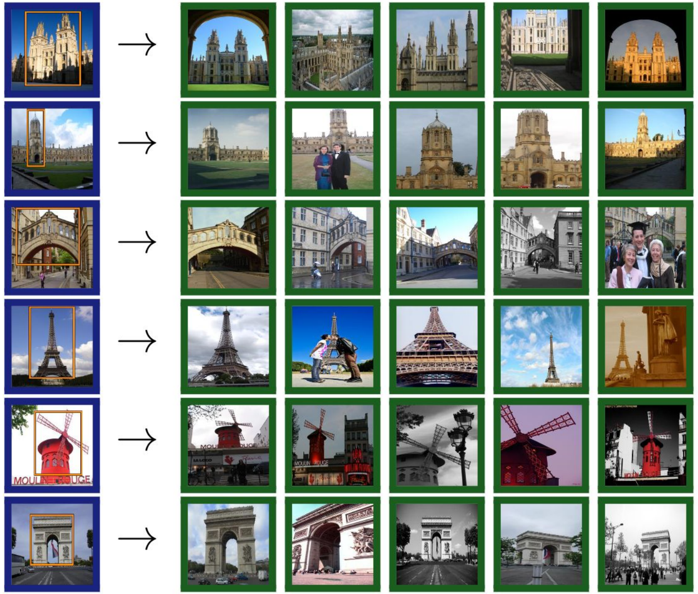


Method	Arch.	Data	Text sup.	kNN val	linear val	linear ReaL	linear V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14_{336}	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	✗	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	✗	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	✗	–	79.8	–	–
MSN	ViT-L/7	INet-1k	✗	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	✗	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	✗	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	✗	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	✗	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	✗	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	✗	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	✗	**83.5**	**86.5**	**89.6**	**78.4**

Experiences & Results

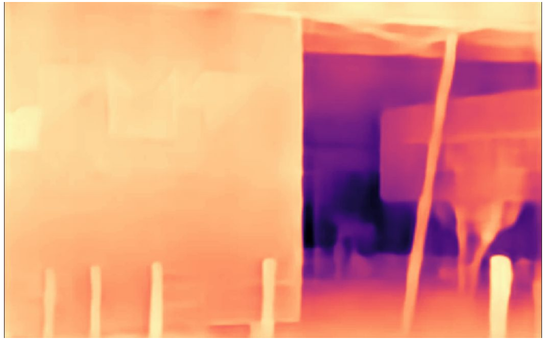
- Instance Recognition
  - Images in database ranked according to the cosine similarities of their features with the ones of a query
  - Outperforms both SSL and weakly supervised methods

		Oxford		Paris		Met			AmsterTime
Feature	Arch	M	H	M	H	GAP	GAP-	ACC	mAP
OpenCLIP	ViT-G/14	50.7	19.7	79.2	60.2	6.5	23.9	34.4	24.6
MAE	ViT-H/14	11.7	2.2	19.9	4.7	7.5	23.5	30.5	4.2
DINO	ViT-B/8	40.1	13.7	65.3	35.3	17.1	37.7	43.9	24.6
iBOT	ViT-L/16	39.0	12.7	70.7	47.0	25.1	54.8	58.2	26.7
DINOv2	ViT-S/14	68.8	43.2	84.6	68.5	29.4	54.3	57.7	43.5
	ViT-B/14	72.9	49.5	90.3	78.6	36.7	63.5	66.1	45.6
	ViT-L/14	**75.1**	**54.0**	**92.7**	**83.5**	**40.0**	68.9	71.6	**50.0**
	ViT-g/14	73.6	52.3	92.1	82.6	36.8	**73.6**	**76.5**	46.7



Experiences & Results

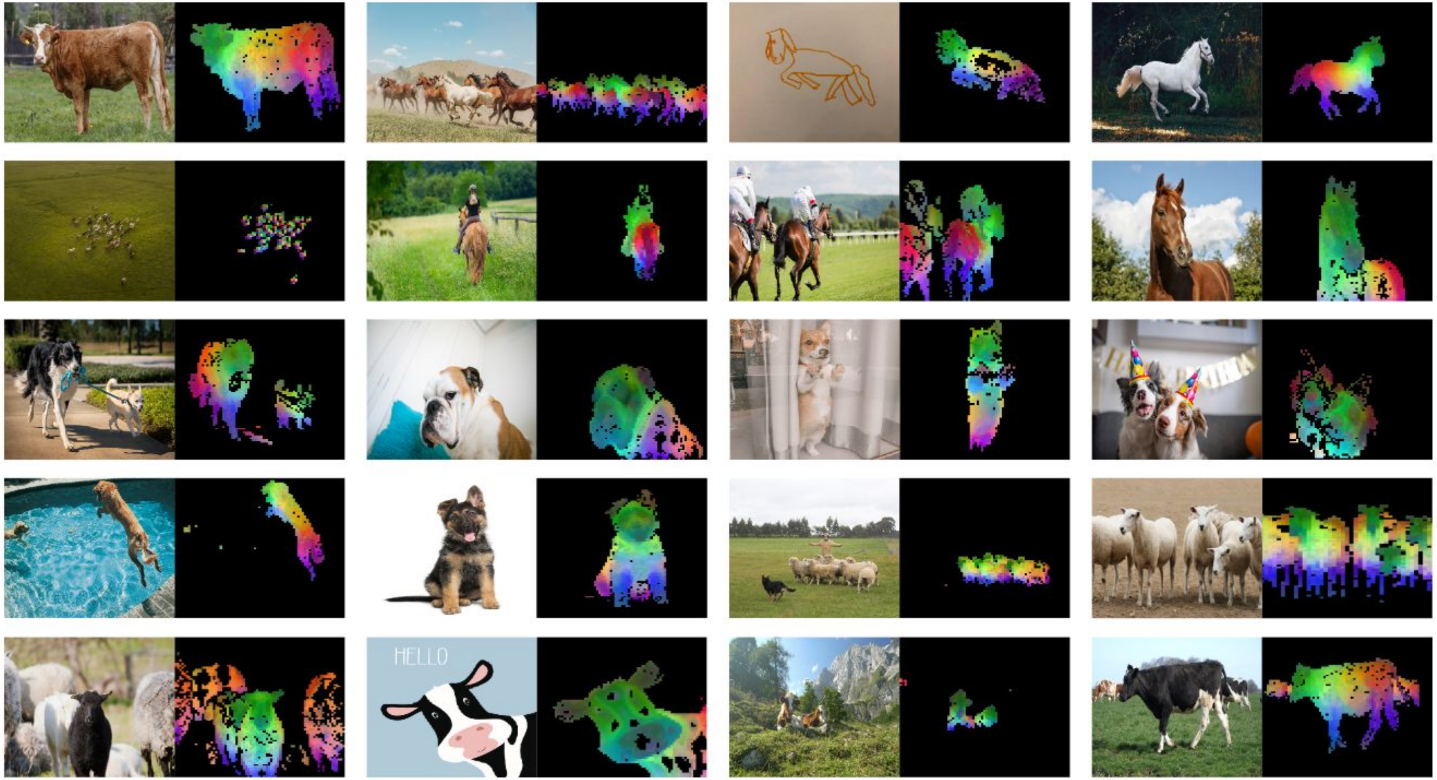
- Semantic segmentation & Depth Estimation



		ADE20k (62.9)		CityScapes (86.9)		Pascal VOC (89.0)	
Method	Arch.	lin.	+ms	lin.	+ms	lin.	+ms
OpenCLIP	ViT-G/14	39.3	46.0	60.3	70.3	71.4	79.2
MAE	ViT-H/14	33.3	30.7	58.4	61.0	67.6	63.3
DINO	ViT-B/8	31.8	35.2	56.9	66.2	66.4	75.6
iBOT	ViT-L/16	44.6	47.5	64.8	74.5	82.3	84.3
DINOv2	ViT-S/14	44.3	47.2	66.6	77.1	81.1	82.6
	ViT-B/14	47.3	51.3	69.4	80.0	82.5	84.9
	ViT-L/14	47.7	**53.1**	70.3	80.9	82.1	86.0
	ViT-g/14	**49.0**	53.0	**71.3**	**81.0**	**83.0**	**86.2**

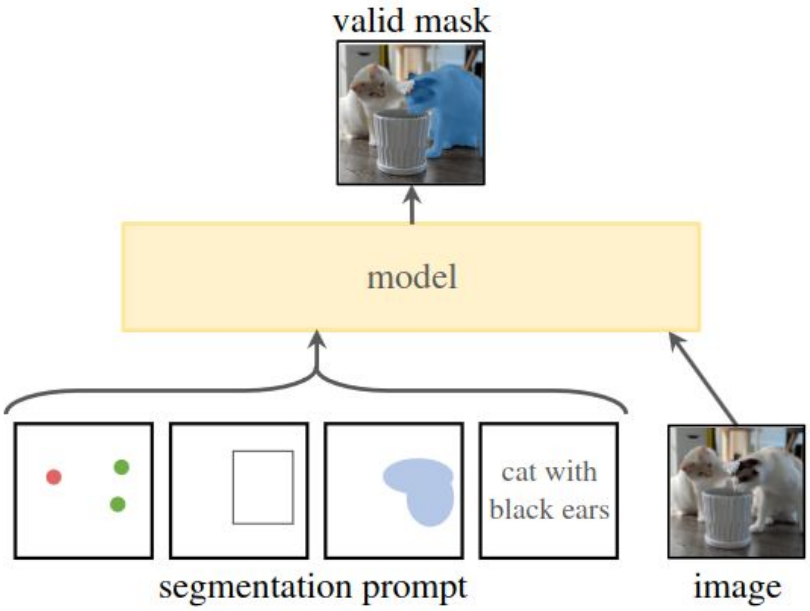
		NYUd (0.330)			KITTI (2.10)			NYUd → SUN RGB-D (0.421)		
Method	Arch.	lin. 1	lin. 4	DPT	lin. 1	lin. 4	DPT	lin. 1	lin. 4	DPT
OpenCLIP	ViT-G/14	0.541	0.510	0.414	3.57	3.21	2.56	0.537	0.476	0.408
MAE	ViT-H/14	0.517	0.483	0.415	3.66	3.26	2.59	0.545	0.523	0.506
DINO	ViT-B/8	0.555	0.539	0.492	3.81	3.56	2.74	0.553	0.541	0.520
iBOT	ViT-L/16	0.417	0.387	0.358	3.31	3.07	2.55	0.447	0.435	0.426
DINOv2	ViT-S/14	0.449	0.417	0.356	3.10	2.86	2.34	0.477	0.431	0.409
	ViT-B/14	0.399	0.362	0.317	2.90	2.59	2.23	0.448	0.400	0.377
	ViT-L/14	0.384	0.333	0.293	2.78	2.50	2.14	0.429	0.396	0.360
	ViT-g/14	**0.344**	**0.298**	**0.279**	**2.62**	**2.35**	**2.11**	**0.402**	**0.362**	**0.338**

Consistent patch mapping



Alternative foundation model for vision: SAM

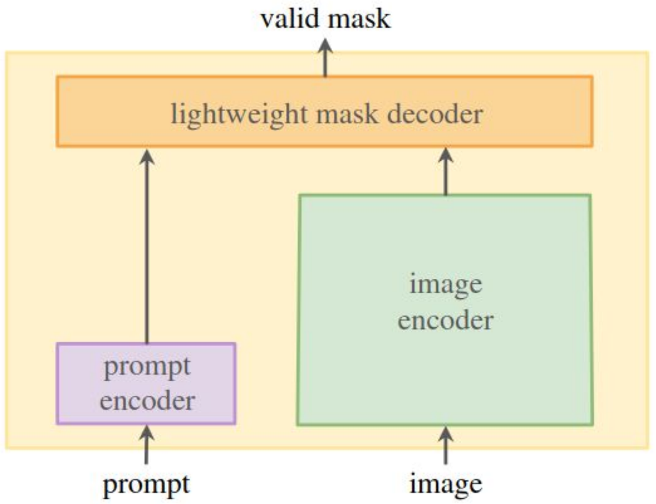
- A generic task
  - Ambition: segment any object
  - Zero- or Few-shot Learning
    - Inspired from NLP
    - Inspired from Hybrid (CLIP)
- Promptable Segmentation
  - Different types
    - Point(s)
    - Boxes
    - Text (?)
  - For training
  - For downstream tasks
  - Return a valid mask for any prompt, even ambiguous
    - Motivations:
      - ability for transfer
      - prompt eng. and composition



(a) **Task**: promptable segmentation

Alternative foundation model for vision: SAM

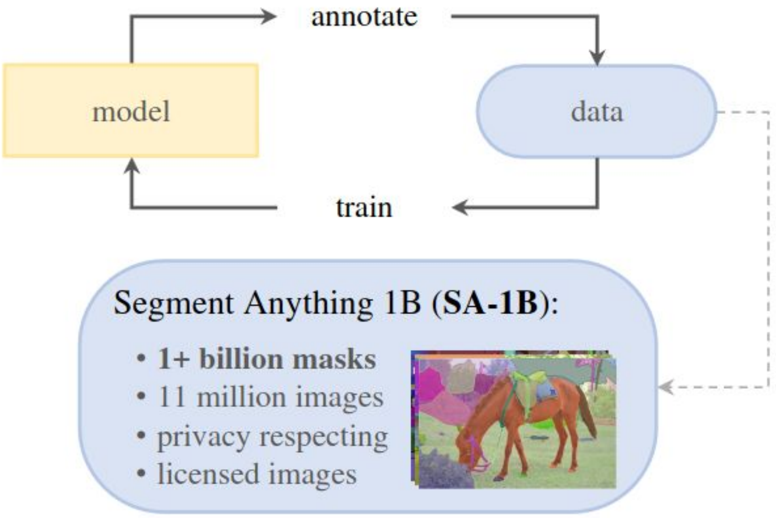
- Architecture
  - Image Encoder (MAE pretrained ViT)
  - Prompt encoder
    - Positional Encoding
    - + Learned Embeddings
  - Lightweight mask decoder
    - Transformer decoder
    - Mask predictor



(b) **Model**: Segment Anything Model (**SAM**)

Alternative foundation model for vision: SAM

- Dataset: SA-1B
  - Assisted Manual Stage
    - Annotators with “brush” / “eraser”
    - “stuff” / “thing”
    - 4.3M masks / 120k images
  - Semi Automatic Stage
    - Display confident masks
    - Ask annotators for additional
    - 5.9M masks / 180k images
  - Fully Automatic Stage
    - 1.1B masks / 11M



(c) **Data**: data engine (top) & dataset (bottom)

Alternative foundation model for vision: SAM

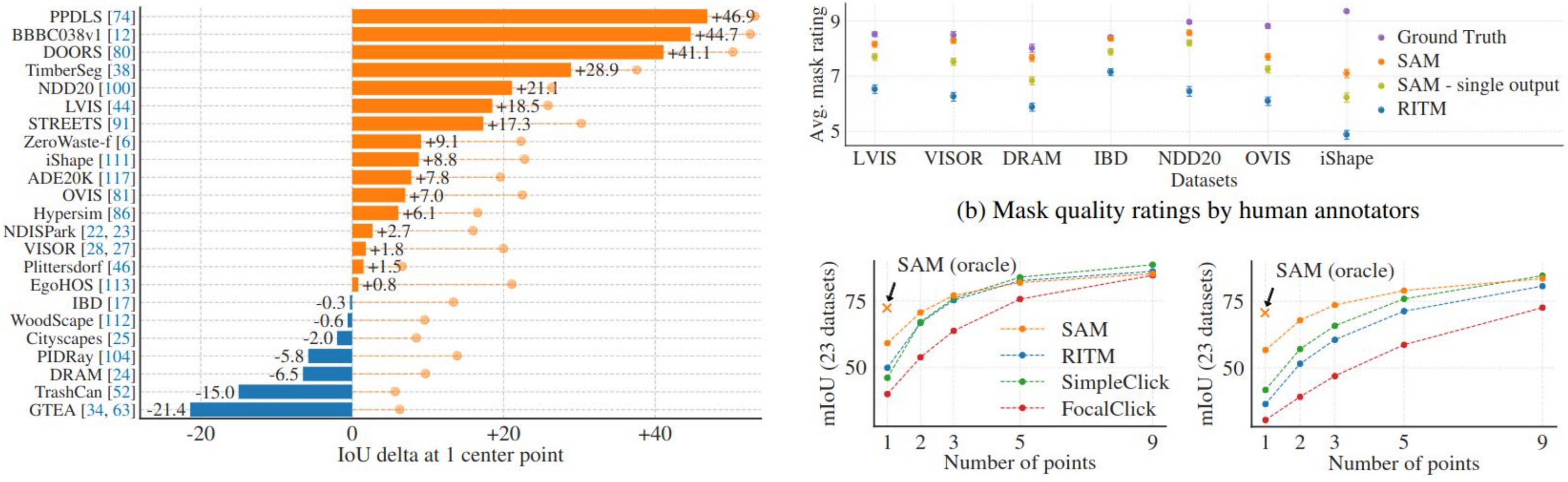
- DINOV2 vs SAM in segmentation : qualitative comparison


IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom
TELECOM
Evolution


IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom
TELECOM
Evolution

Evaluation tasks

- Zero-Shot Single Point Valid Mask Evaluation
    - IoU
    - Quality rating by human annotators



(a) SAM *vs*. RITM [92] on 23 datasets

(b) Mask quality ratings by human annotators

(c) Center points (default)

(d) Random points

Evaluation tasks

- Zero-Shot Edge Detection
  - Segment 16x16 points
  - NMS ; Sobel Filter ; Edge NMS

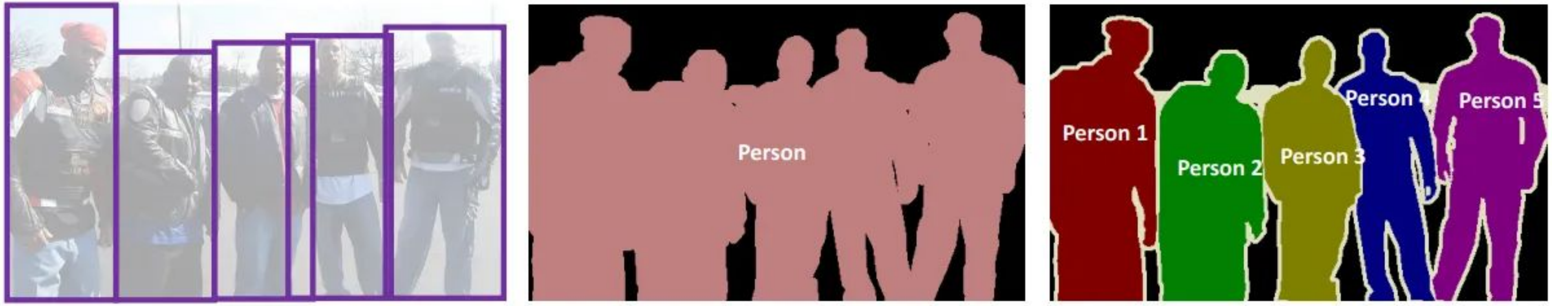
method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Table 3: Zero-shot transfer to edge detection on BSDS500.

Evaluation tasks

- Zero-Shot Instance Segmentation
  - Instance segmentation
  - Running a detector, applying sam on boxes

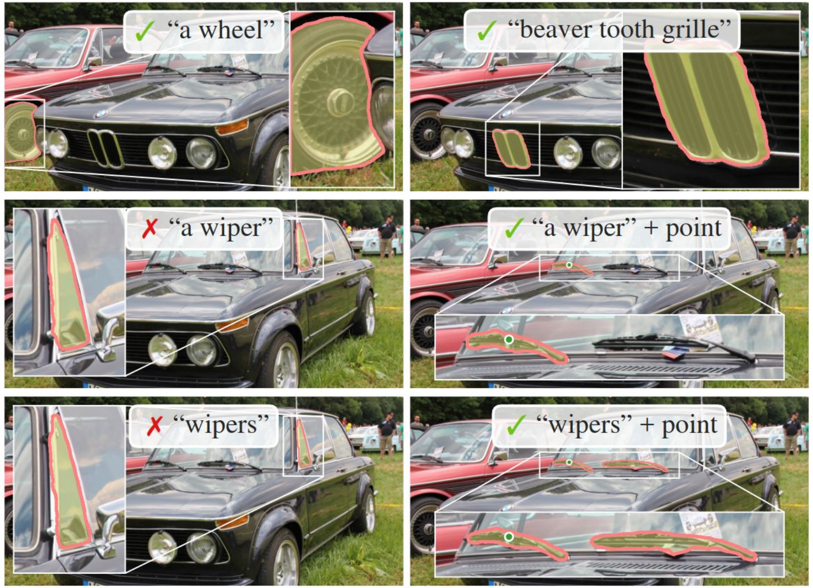
method	COCO [66] AP	AP^S	AP^M	AP^L	LVIS v1 [44] AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5



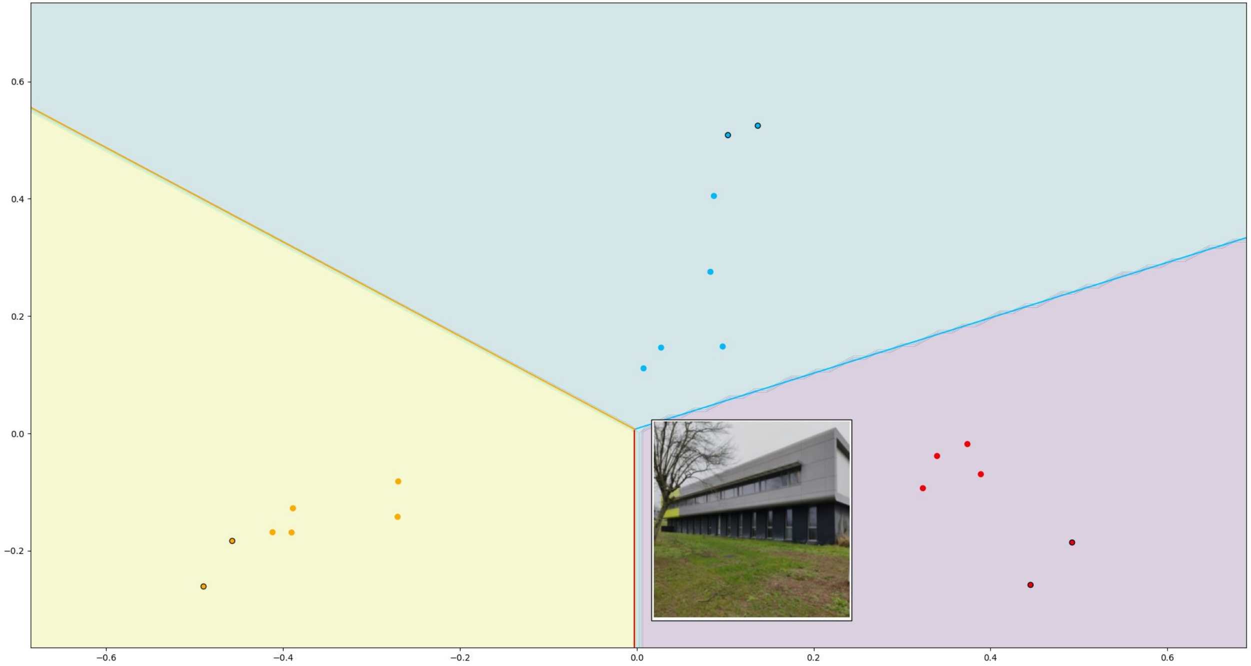
Object Detection Semantic Segmentation **Instance Segmentation**

Evaluation tasks

- Zero-Shot Text-to-Mask
  - Instance segmentation
  - Running a detector, applying sam on boxes



Hands on : few shot classification with DINOv2

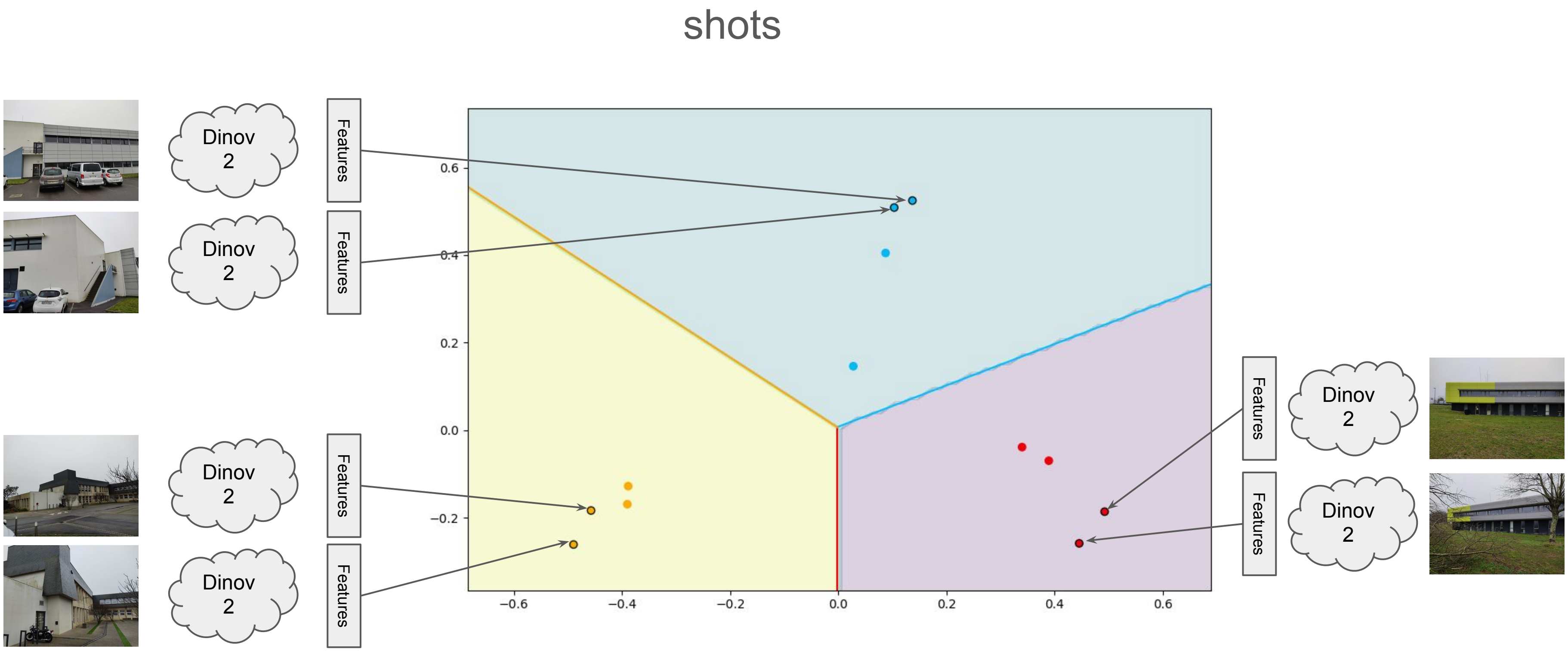


Hands on : few shot classification with DINOv2

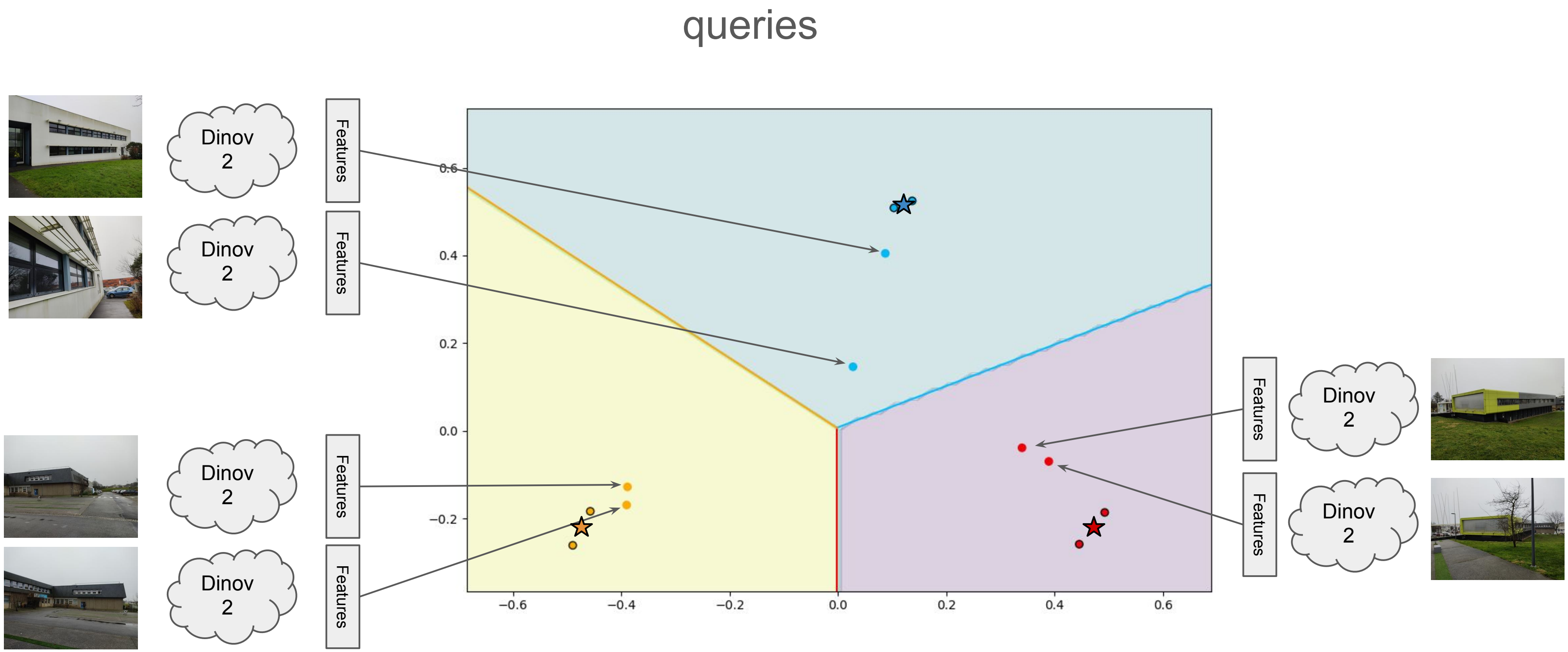
shots

queries

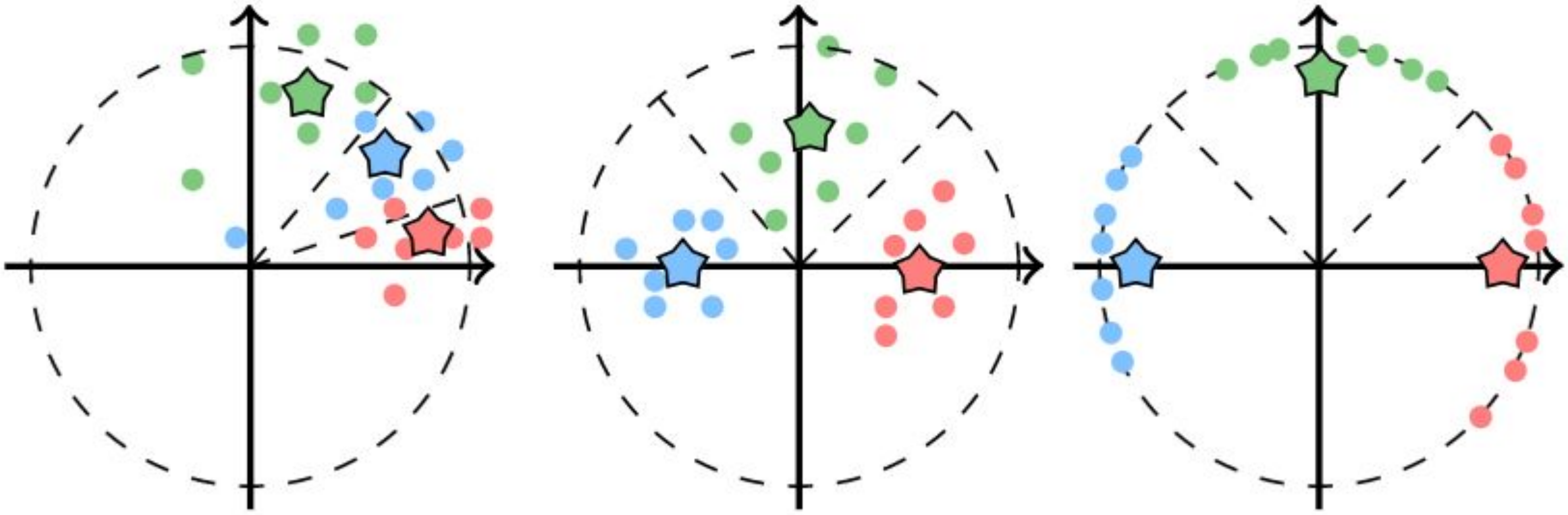
Hands on : few shot classification with DINOv2



Hands on : few shot classification with DINOv2



Center and project



Center and project