NEURAL INFORMATION
PROCESSING SYSTEMS

DON’T NEED RETRAINING: A Mixture of DETR and Vision Foundation Models for Cross-Domain Few-Shot Object Detection

Changhan Liu, Xunzhi Xiang, Zixuan Duan, Wenbin Li, Qi Fan, Yang Gao

School of Intelligence Science and Technology, Nanjing University, China

Background

Cross Domain Few Shot Object Detection

- **Definition :**

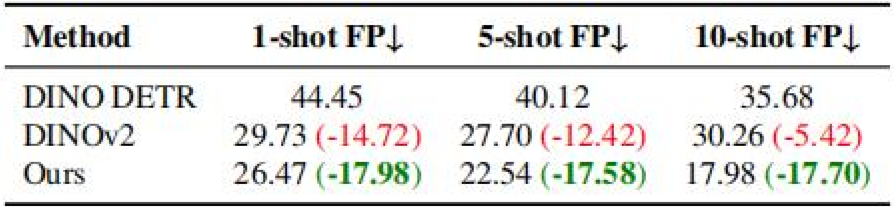
  CrossDomain Few-Shot Object Detection (CD-FSOD) aims to generalize object detection models to detect novel classes in unseen domains by using a few training samples.

This challenging task typically requires model to combine strong generalization and accurate localization capability

Motivation

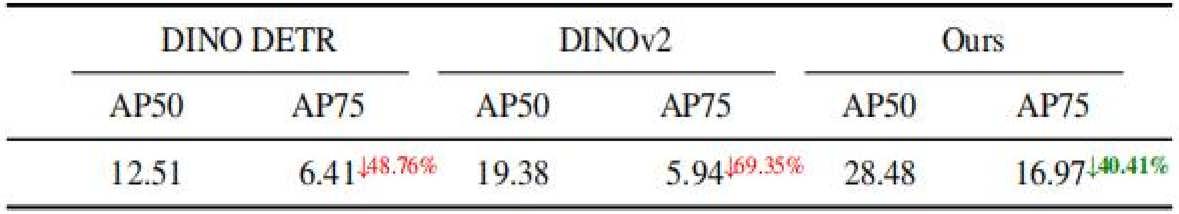
- Existing well-trained detectors typically have strong localization capabilities but lack generalization.
- Vision foundation models (VFMs) generally exhibit better generalization but lack accurate localization capabilities.

Comparison Results of Generalization



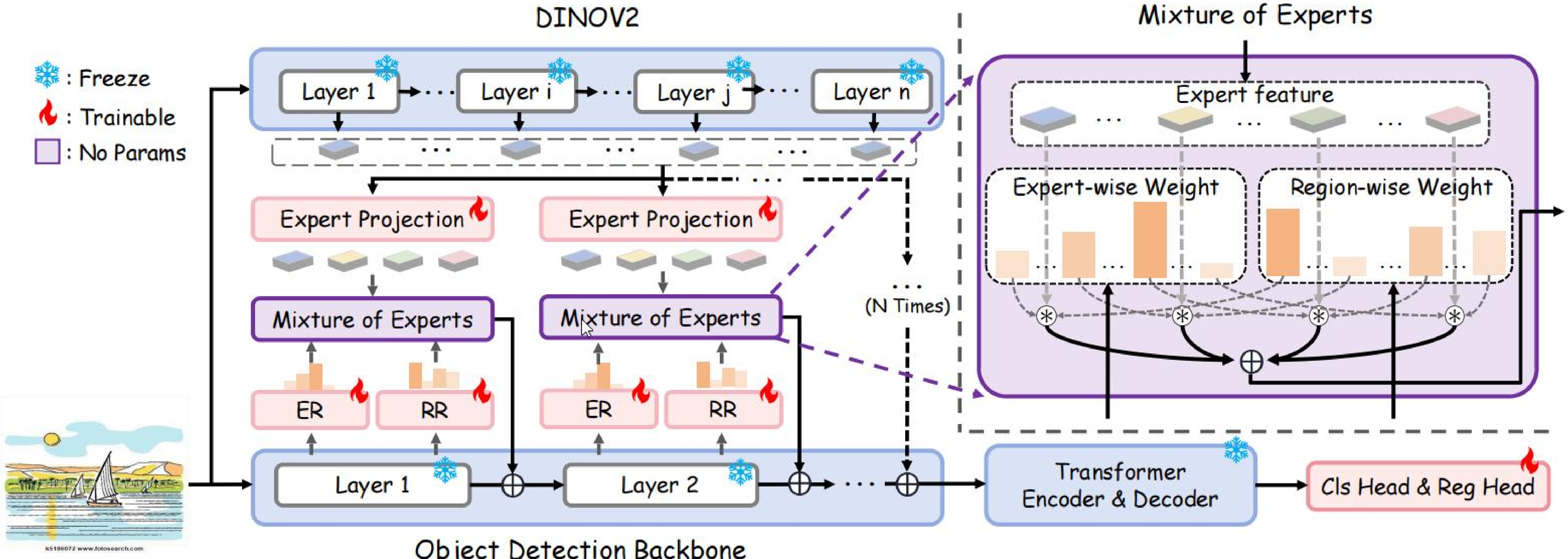
Method	1-shot FP↓	5-shot FP↓	10-shot FP↓
DINO DETR	44.45	40.12	35.68
DINOv2	29.73 (-14.72)	27.70 (-12.42)	30.26 (-5.42)
Ours	26.47 (**-17.98**)	22.54 (**-17.58**)	17.98 (**-17.70**)

Comparison Results of Localization Capability



DINO DETR		DINOv2		Ours	
AP50	AP75	AP50	AP75	AP50	AP75
12.51	6.41$^{\downarrow 48.76\%}$	19.38	5.94$^{\downarrow 69.35\%}$	28.48	16.97$^{\downarrow \mathbf{40.41\%}}$

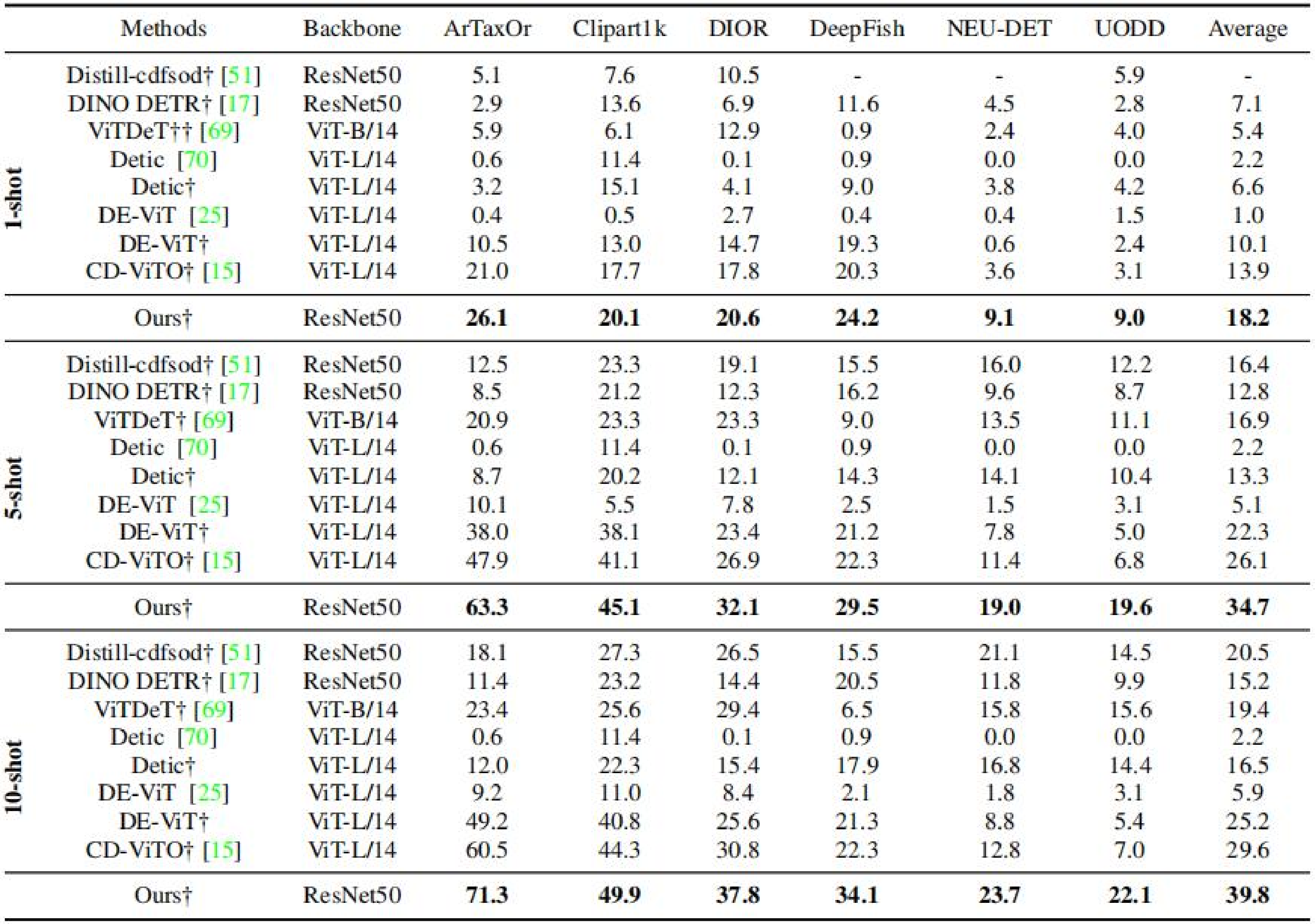
Method



- **Expert-wise Router (ER) and Region-wise Router (RR) :** The ER module generates expert-wise gating weight to select the appropriate VFM expert features for the detector features at different layers. The RR module generates region-wise gating weight to filter out the invalid background regions in the VFM feature map
- **Shared Expert Projection (SEP) and Private Expert Projection (PEP) :** The SEP module projects the shared image feature contained in different expert features. The PEP module projects the private image feature contained in each expert feature.

Experiments

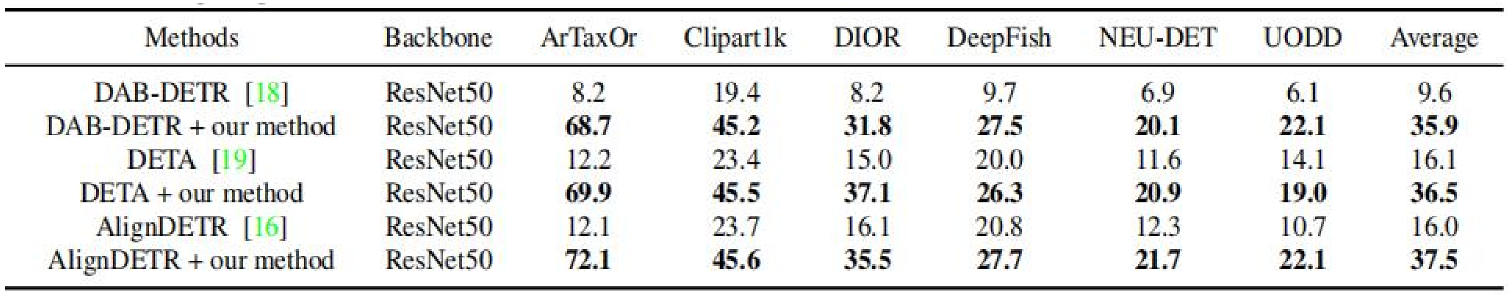
Comparison with State-of-The-Arts



	Methods	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Average
1-shot	Distill-cdfsod† [51]	ResNet50	5.1	7.6	10.5	-	-	5.9	-
	DINO DETR† [17]	ResNet50	2.9	13.6	6.9	11.6	4.5	2.8	7.1
	ViTDeT†† [69]	ViT-B/14	5.9	6.1	12.9	0.9	2.4	4.0	5.4
	Detic [70]	ViT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic†	ViT-L/14	3.2	15.1	4.1	9.0	3.8	4.2	6.6
	DE-ViT [25]	ViT-L/14	0.4	0.5	2.7	0.4	0.4	1.5	1.0
	DE-ViT†	ViT-L/14	10.5	13.0	14.7	19.3	0.6	2.4	10.1
	CD-ViTO† [15]	ViT-L/14	21.0	17.7	17.8	20.3	3.6	3.1	13.9
	Ours†	ResNet50	**26.1**	**20.1**	**20.6**	**24.2**	**9.1**	**9.0**	**18.2**
5-shot	Distill-cdfsod† [51]	ResNet50	12.5	23.3	19.1	15.5	16.0	12.2	16.4
	DINO DETR† [17]	ResNet50	8.5	21.2	12.3	16.2	9.6	8.7	12.8
	ViTDeT† [69]	ViT-B/14	20.9	23.3	23.3	9.0	13.5	11.1	16.9
	Detic [70]	ViT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic†	ViT-L/14	8.7	20.2	12.1	14.3	14.1	10.4	13.3
	DE-ViT [25]	ViT-L/14	10.1	5.5	7.8	2.5	1.5	3.1	5.1
	DE-ViT†	ViT-L/14	38.0	38.1	23.4	21.2	7.8	5.0	22.3
	CD-ViTO† [15]	ViT-L/14	47.9	41.1	26.9	22.3	11.4	6.8	26.1
	Ours†	ResNet50	**63.3**	**45.1**	**32.1**	**29.5**	**19.0**	**19.6**	**34.7**
10-shot	Distill-cdfsod† [51]	ResNet50	18.1	27.3	26.5	15.5	21.1	14.5	20.5
	DINO DETR† [17]	ResNet50	11.4	23.2	14.4	20.5	11.8	9.9	15.2
	ViTDeT† [69]	ViT-B/14	23.4	25.6	29.4	6.5	15.8	15.6	19.4
	Detic [70]	ViT-L/14	0.6	11.4	0.1	0.9	0.0	0.0	2.2
	Detic†	ViT-L/14	12.0	22.3	15.4	17.9	16.8	14.4	16.5
	DE-ViT [25]	ViT-L/14	9.2	11.0	8.4	2.1	1.8	3.1	5.9
	DE-ViT†	ViT-L/14	49.2	40.8	25.6	21.3	8.8	5.4	25.2
	CD-ViTO† [15]	ViT-L/14	60.5	44.3	30.8	22.3	12.8	7.0	29.6
	Ours†	ResNet50	**71.3**	**49.9**	**37.8**	**34.1**	**23.7**	**22.1**	**39.8**

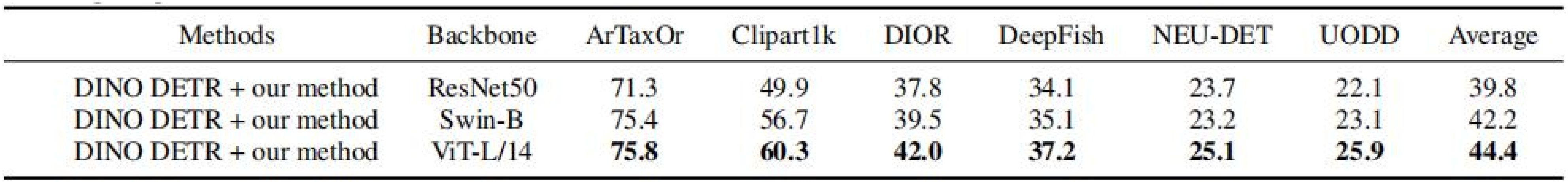
Experiments

Method Extensibility Performance



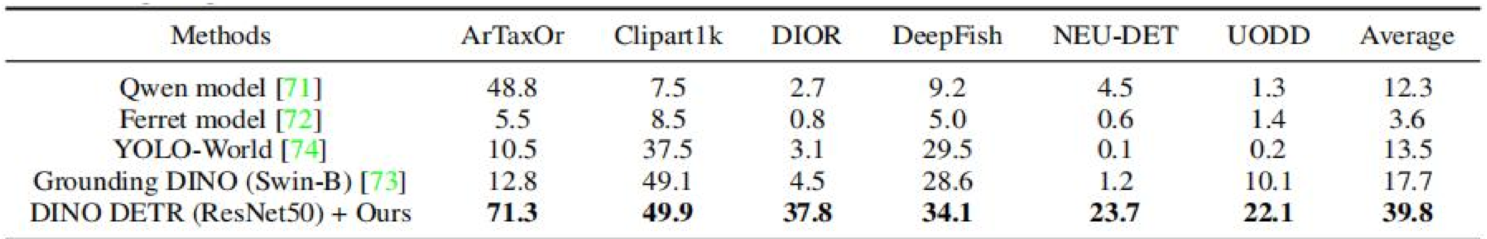
Methods	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Average
DAB-DETR [18]	ResNet50	8.2	19.4	8.2	9.7	6.9	6.1	9.6
DAB-DETR + our method	ResNet50	**68.7**	**45.2**	**31.8**	**27.5**	**20.1**	**22.1**	**35.9**
DETA [19]	ResNet50	12.2	23.4	15.0	20.0	11.6	14.1	16.1
DETA + our method	ResNet50	**69.9**	**45.5**	**37.1**	**26.3**	**20.9**	**19.0**	**36.5**
AlignDETR [16]	ResNet50	12.1	23.7	16.1	20.8	12.3	10.7	16.0
AlignDETR + our method	ResNet50	**72.1**	**45.6**	**35.5**	**27.7**	**21.7**	**22.1**	**37.5**

Method Performance on Different Backbones



Methods	Backbone	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Average
DINO DETR + our method	ResNet50	71.3	49.9	37.8	34.1	23.7	22.1	39.8
DINO DETR + our method	Swin-B	75.4	56.7	39.5	35.1	23.2	23.1	42.2
DINO DETR + our method	ViT-L/14	**75.8**	**60.3**	**42.0**	**37.2**	**25.1**	**25.9**	**44.4**

Comparison with MLLMs and OVMs



Methods	ArTaxOr	Clipart1k	DIOR	DeepFish	NEU-DET	UODD	Average
Qwen model [71]	48.8	7.5	2.7	9.2	4.5	1.3	12.3
Ferret model [72]	5.5	8.5	0.8	5.0	0.6	1.4	3.6
YOLO-World [74]	10.5	37.5	3.1	29.5	0.1	0.2	13.5
Grounding DINO (Swin-B) [73]	12.8	49.1	4.5	28.6	1.2	10.1	17.7
DINO DETR (ResNet50) + Ours	**71.3**	**49.9**	**37.8**	**34.1**	**23.7**	**22.1**	**39.8**

Thank you !