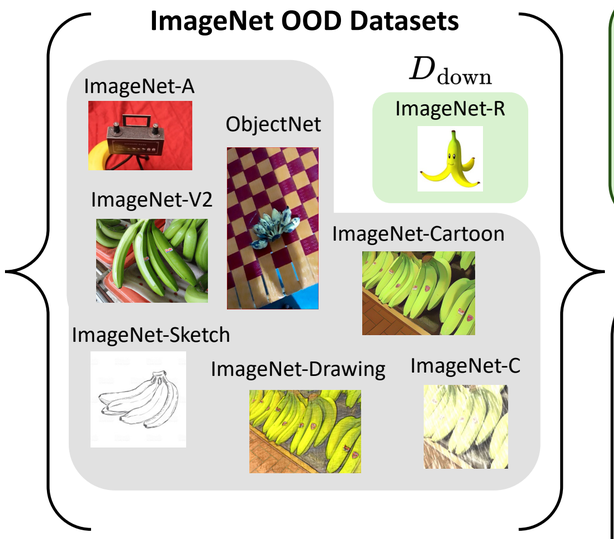
ImageNet-RIB Benchmark: Large Pre-Training Datasets Don't Guarantee Robustness after Fine-Tuning

Jaedong Hwang Brian Cheung Zhang-Wei Hong Akhilan Boopathy Pulkit Agrawal Ila R Fiete

NEURAL INFORMATION PROCESSING SYSTEMS

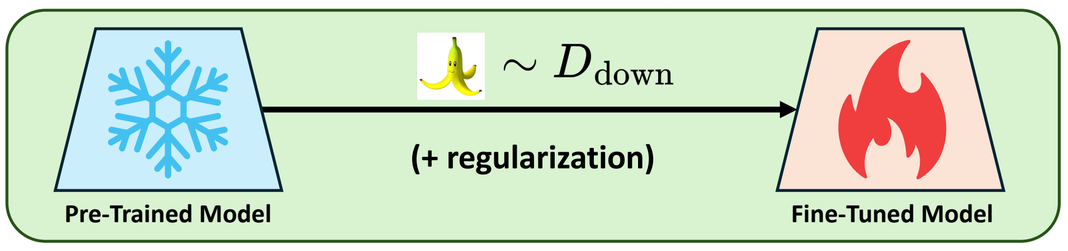
❑ ImageNet-RIB (Robustness Improvement Benchmark)

1. Choose one dataset



4. Repeat 1-3

2. Fine-tune on the downstream dataset



3. Evaluate on other OOD datasets

	IN-V2	IN-A	IN-Sketch	ObjNet	IN-Cartoon	IN-Drawing	IN-C
Acc(❄)	66	15	28	26	66	39	56
Acc(🔥)	59	21	47	32	61	51	52
Robustness Change	-5	+6	+19	+6	-5	+12	-4

▪ Metrics

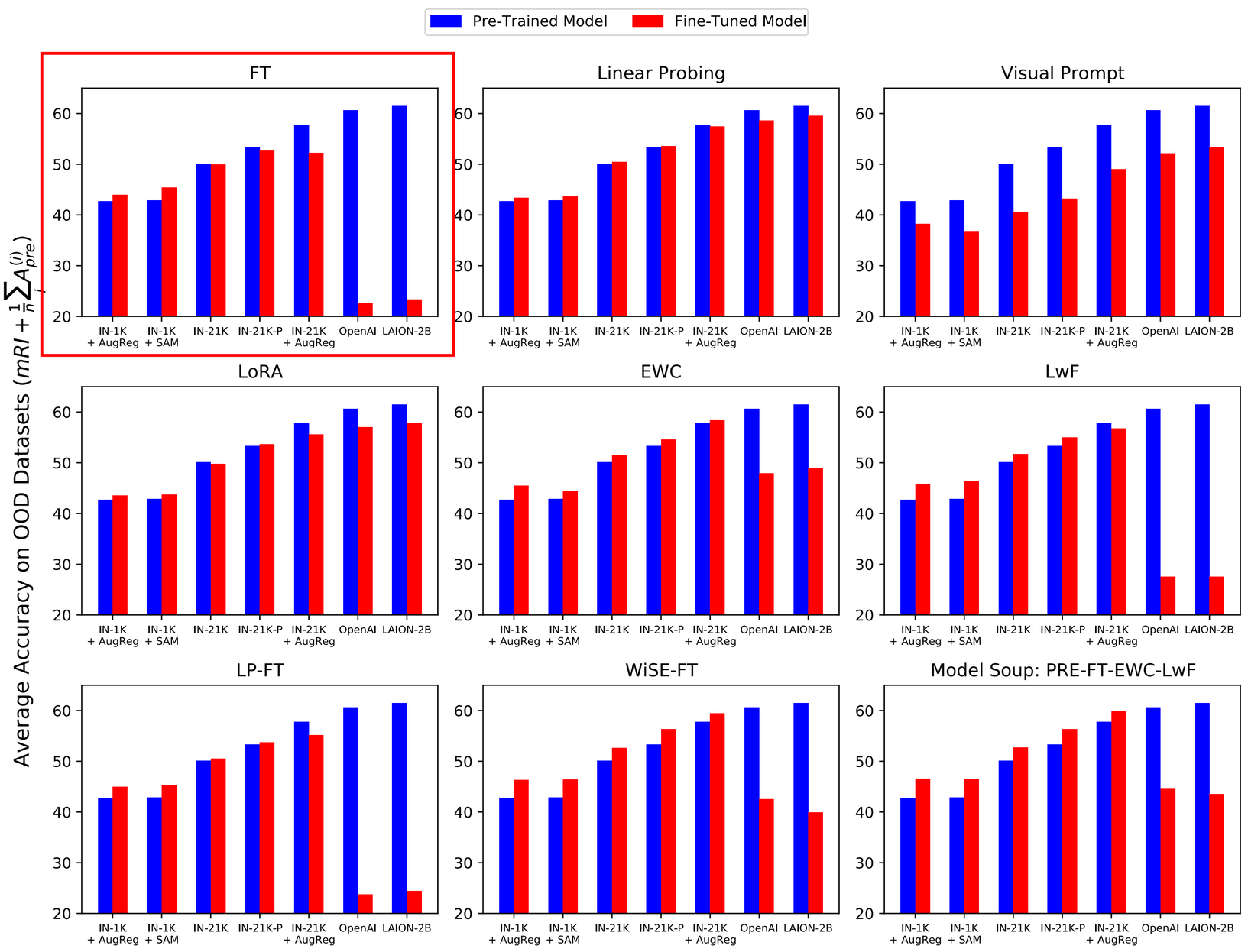
▪ **Robustness Improvement (RI):**
Average Accuracy Difference between fine-tuned model and pre-trained model on OOD datasets

$$RI_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} A_i^{(j)} - A_{\text{pre}}^{(j)}$$

$$mRI = \sum_i^n RI_i$$

Models Pre-Trained on Richer and Larger Datasets Are Worse on OOD Datasets after Fine-Tuning

E. Munch. 'The scream.' Wellcome Library

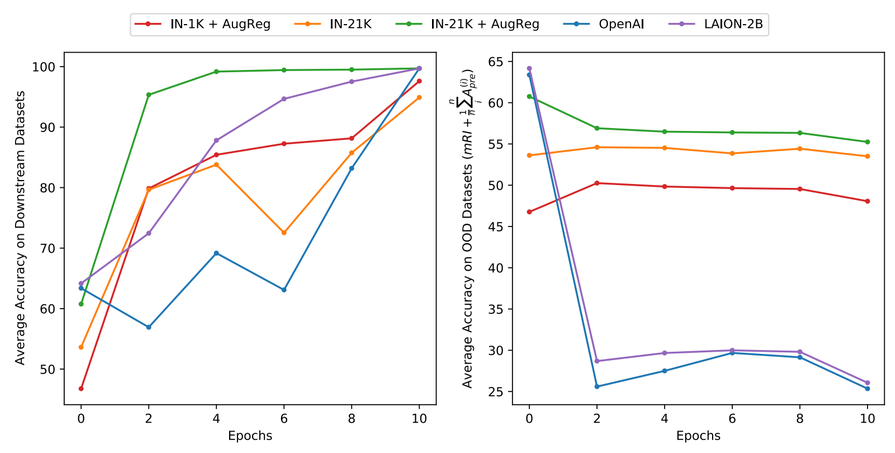


▪ Accuracy on each OOD dataset after fine-tuning on D_{down}

Method	Downstream Dataset	D_{pre} IN	Realistic OOD IN-V2	IN-A	IN-R	IN-Sketch	ObjNet	Synthetic OOD IN-Cartoon	IN-Drawing	IN-C
Pre-Trained		79.2	66.4	15.0	38.0	28.0	25.7	66.2	39.1	56.0
FT	IN-V2	78.4	-	**25.2**	41.9	29.2	37.1	64.7	40.4	57.4
	IN-A	72.9	60.6	-	36.7	24.9	35.0	55.3	32.6	53.5
	IN-R	69.8	59.2	20.9	-	**46.7**	32.0	61.3	51.4	52.0
	IN-Sketch	75.7	63.9	17.3	**59.1**	-	33.0	66.3	50.8	53.8
	ObjNet	74.4	62.2	24.9	36.3	25.1	-	55.6	33.6	52.3
	IN-Cartoon	85.2	63.5	19.9	40.5	29.5	33.5	-	41.2	51.3
	IN-Drawing	81.5	62.9	16.5	41.1	32.7	32.4	64.2	-	56.0
	IN-C	**99.8**	61.1	13.9	37.0	25.1	27.7	**92.2**	**70.2**	-
Model Soup PRE-FT-EWC-LwF (Wortsman et al., 2022a)	IN-V2	79.8	-	21.0	41.0	29.7	36.0	66.9	41.7	58.0
	IN-A	78.3	66.4	-	39.7	28.5	37.5	63.7	38.4	57.8
	IN-R	78.9	67.1	23.1	-	**45.9**	37.2	69.6	55.8	59.6
	IN-Sketch	78.9	66.6	17.5	54.0	-	34.6	69.1	49.8	57.5
	ObjNet	79.3	67.4	24.1	40.3	29.1	-	64.9	40.6	57.7
	IN-Cartoon	83.7	66.4	18.9	41.8	30.6	34.7	-	43.6	56.2
	IN-Drawing	82.6	66.9	18.4	43.0	34.0	35.2	68.7	-	**59.7**
	IN-C	92.6	67.5	18.6	42.3	30.6	35.3	81.3	57.3	-

▪ Accuracy on Downstream Dataset and OOD datasets

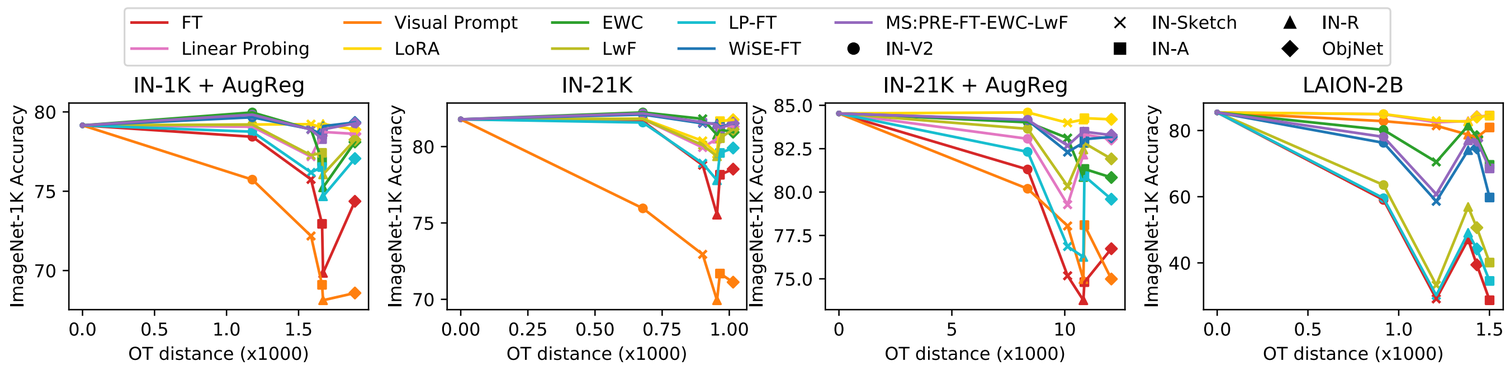
▪ Severe robustness degradation is not due to overfitting



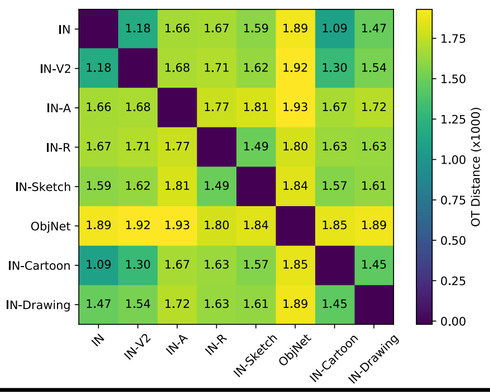
▪ **Continual Learning with Post-Hoc Robust Fine-Tuning Method [1] Perform Best**

❑ Optimal Transport Dataset Distance [2] Aligns with ImageNet-1K Accuracy Drop During Fine-Tuning

▪ **IN-1K Accuracy after Fine-Tuning on Each Downstream Dataset**



▪ **OTDD on Feature Space from IN-1K + AugReg Pre-Trained ViT-B/16**



▪ It matches with design principle of each datasets

❑ **Reference**
[1] Wortsman, Mitchell, et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time." ICML. 2022.
[2] Alvarez-Melis, David, and Nicolo Fusi. "Geometric dataset distances via optimal transport." NeurIPS. 2020.