Mixture of Adversarial LoRAs: Boosting Robust Generalization in Meta-Tuning

Xu Yang
City University of Hong Kong

Chen Liu*
City University of Hong Kong

Ying Wei*
Zhejiang University

*Corresponding authors

Outline

Background

Robust Generalization

Motivation

Proposed Methods

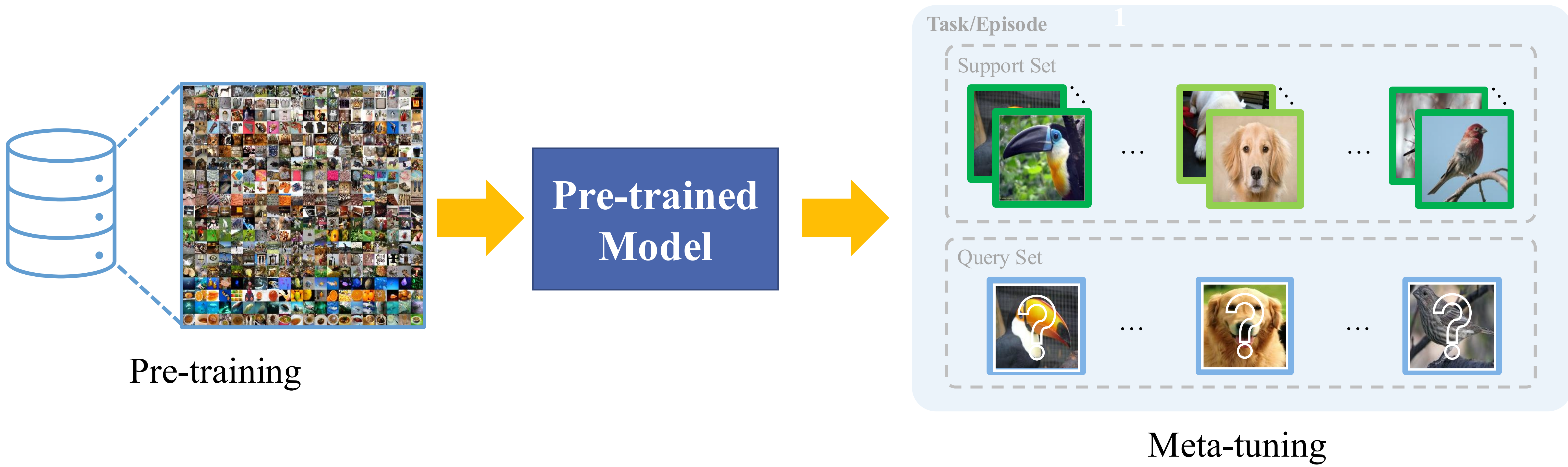
Framework

Experiments

Analysis

Meta-Tuning in Large-scale Pre-trained Models

- Large-scale pre-trained vision transformers have revolutionized the few-shot learning area [1]
- Meta-tuning equips pre-trained models with quick adaptation capability by training on a handful of few-shot tasks



[1] Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In CVPR, 2022

Does Meta-Tuned Models Generalize Well?

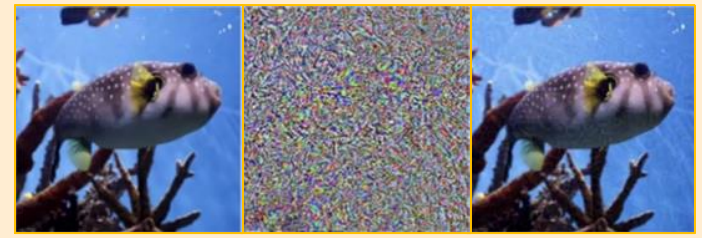
In-Distribution

- **Concept shift:** training and test samples are collected from the same environment yet from mutually exclusive classes

Out-of-Distribution

- **Domains** of images (e.g., from IN to QuickDraw) or **granularity** of categories (e.g., from iNaturalist to Plant Disease) in unseen tasks deviate from those in the training tasks.

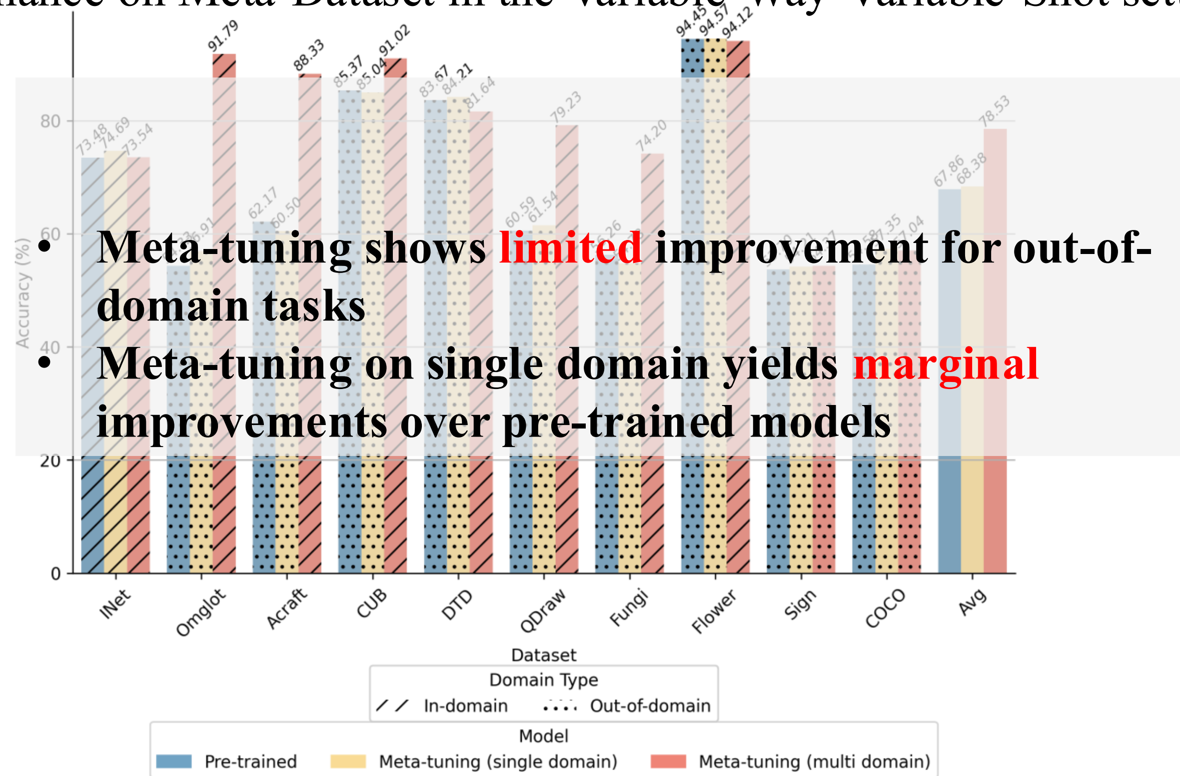
Robustness



Puffer, 97.99% crab, 100%

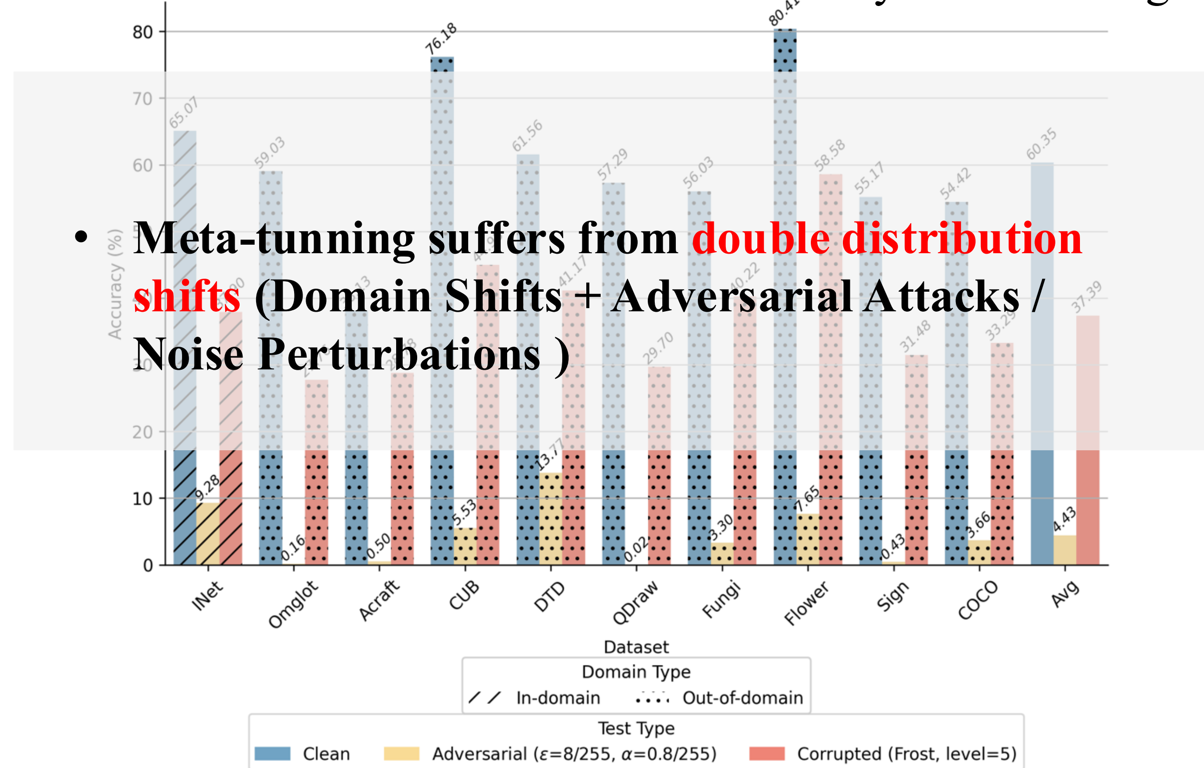
- **Adversarial Vulnerability:** human-imperceptible perturbations
- **Visual corruptions:** weather, noise, blur, etc.

Performance on Meta-Dataset in the Variable-Way-Variable-Shot setting



- **Meta-tuning shows limited improvement for out-of-domain tasks**
- **Meta-tuning on single domain yields marginal improvements over pre-trained models**

Performance on Meta-Dataset in the 5-way 1-shot setting



- **Meta-tunning suffers from double distribution shifts (Domain Shifts + Adversarial Attacks / Noise Perturbations)**

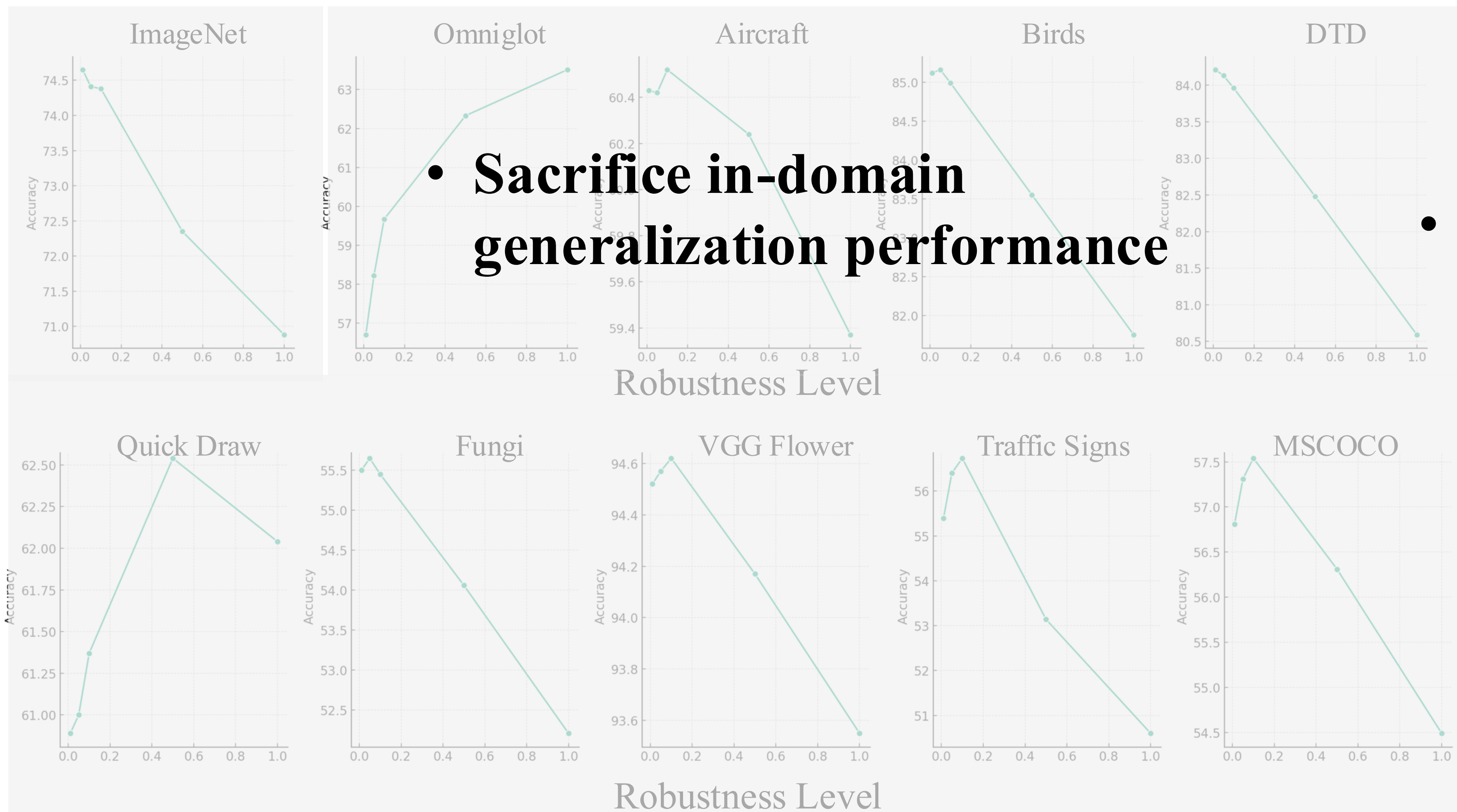
What is Relationship between Adversarial Robustness & Distribution Shift?

- Meta-train on ImageNet using adversarial examples generated under different perturbation budgets ϵ
- Meta-test on in-domain and out-of-domain datasets

$$\min_{\theta} \max_{\|\delta\|_{\infty} \le \epsilon} \mathcal{L}(f_{\theta}(x+\delta), y)$$

$$\delta \leftarrow \Pi_{\epsilon}\left(\delta + \alpha \cdot \mathrm{sign}\left(\nabla_{\delta} \mathcal{L}(f_{\theta}(x+\delta), y)\right)\right)$$

f_{θ}: classifier with parameters θ
$\mathcal{L}$: CE loss
δ: perturbations
ϵ: robustness level
α: step size
Π_{ϵ}: projection for constraints



- **Sacrifice in-domain generalization performance**

- **OOD performance can be improved with suitable robustness levels**

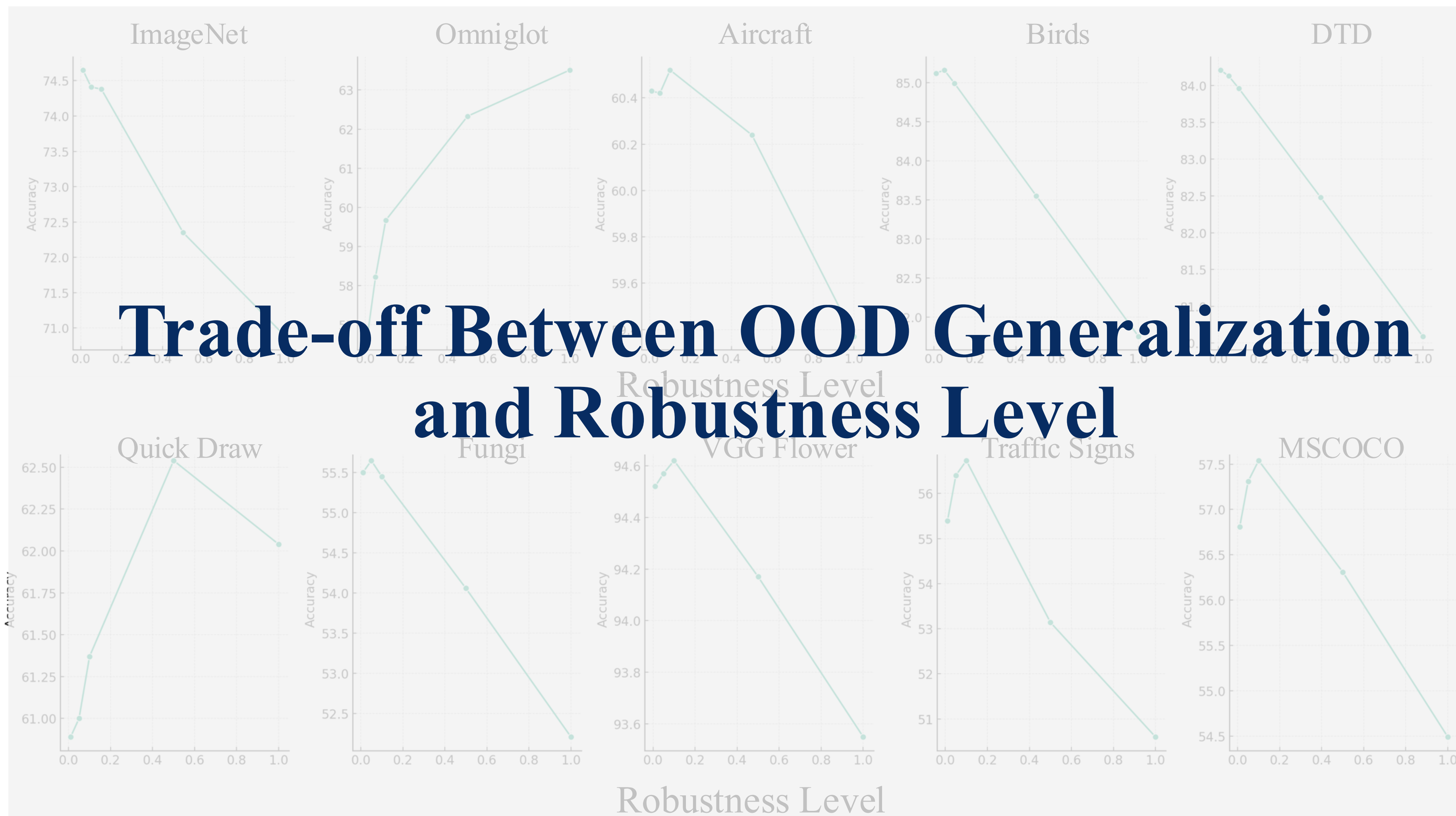
What is Relationship between Adversarial Robustness & Distribution Shift?

- Meta-train on ImageNet using adversarial examples generated under different perturbation budgets $\boldsymbol{\epsilon}$
- Meta-test on OOD datasets

$$\min_{\theta} \max_{\|\delta\|_\infty \le \epsilon} \mathcal{L}(f_\theta(x+\delta), y)$$

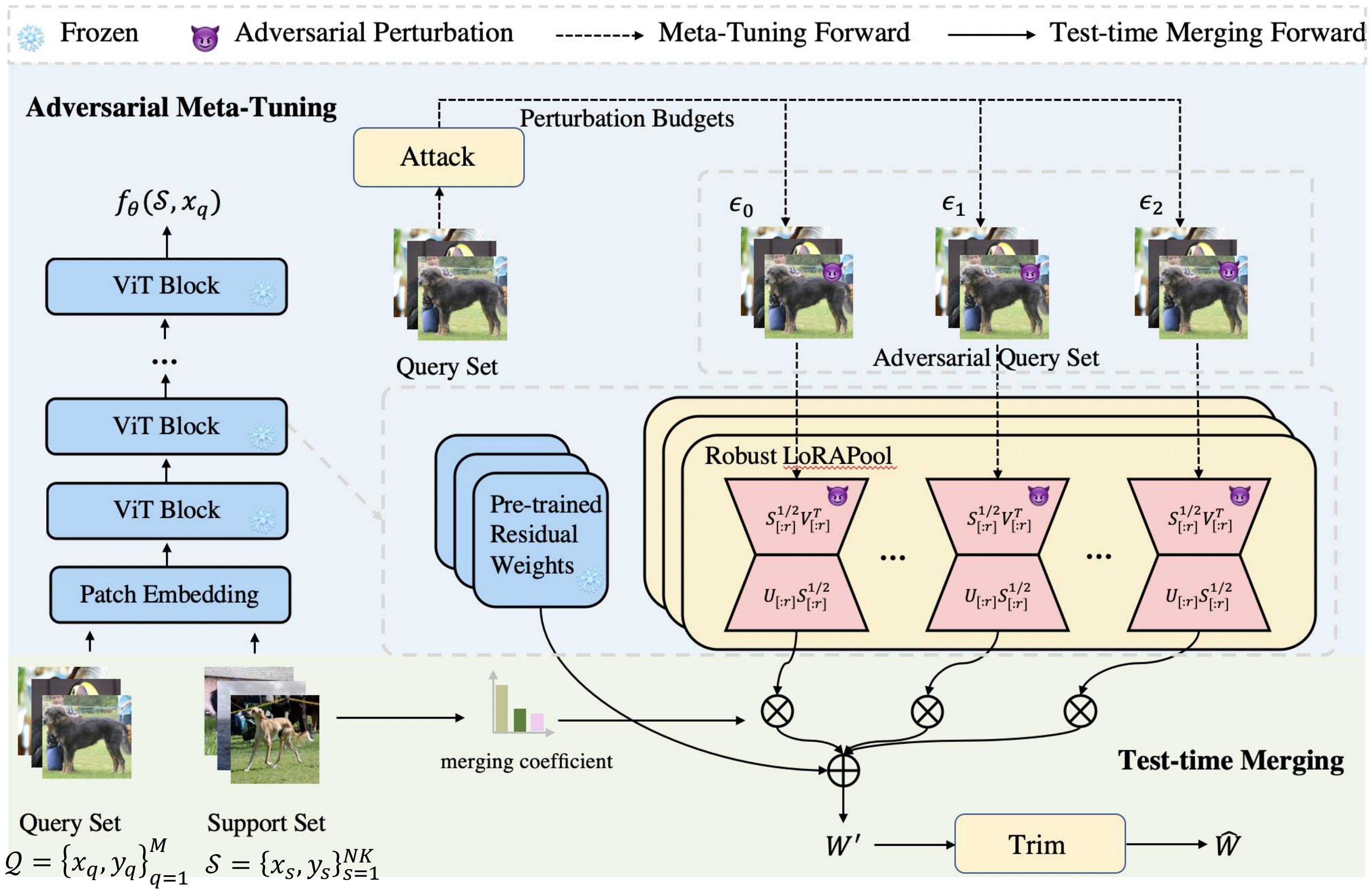
$$\delta \leftarrow \Pi_\epsilon\left(\delta + \alpha \cdot \mathrm{sign}\left(\nabla_\delta \mathcal{L}(f_\theta(x+\delta), y)\right)\right)$$

f_θ: classifier with parameters θ
$\mathcal{L}$: CE loss
δ: perturbations
ϵ: robustness level
α: step size
Π_ϵ: projection for constraints



Trade-off Between OOD Generalization and Robustness Level

Adaptive Robust LoRAPool



Training Objective

$$\mathcal{L}_{clean} = \mathcal{L}_{CE}\left(f_{W^{res}+AB}(\mathcal{S}, x_q), y_q\right)$$ *(Clean Cross-Entropy loss)*

$$\mathcal{L}_{adv} = D_{\mathrm{KL}}\left(f_{W^{res}+AB}(\mathcal{S}, x_q^{adv}) \parallel f_{W^{res}+AB}(\mathcal{S}, x_q)\right)$$

$$\mathcal{L} = \mathcal{L}_{clean} + \lambda_{adv}\mathcal{L}_{adv}$$ *(Adversarial Kullback-Leibler divergence loss)*

Generate Adversarial Query Set

$$\max_{\|\delta\|_\infty \le \epsilon_i} \mathcal{L}\left(f_\theta(\mathcal{S}, x_q + \delta), y_q\right)$$

$$\delta \leftarrow \Pi_{\epsilon_i}\left(\delta + \alpha \cdot \mathrm{sign}\left(\nabla_\delta \mathcal{L}\left(f_\theta(\mathcal{S}, x_q + \delta), y_q\right)\right)\right)$$

(Sample the i-th attack configuration candidates to generate adversarial query sets with different robustness strength)

Adversarial Perturbation on Singular Values and Vectors

$$A = U_{[:r]}\mathrm{diag}\left(S_{[:r]}^{1/2}\right)$$

$$\mathrm{B} = \mathrm{diag}\left(S_{[:r]}^{1/2}\right) V_{[:r]}^T$$

$$W^{\mathrm{res}} = U_{[r:]}\mathrm{diag}(S_{[r:]})V_{[r:]}^T$$

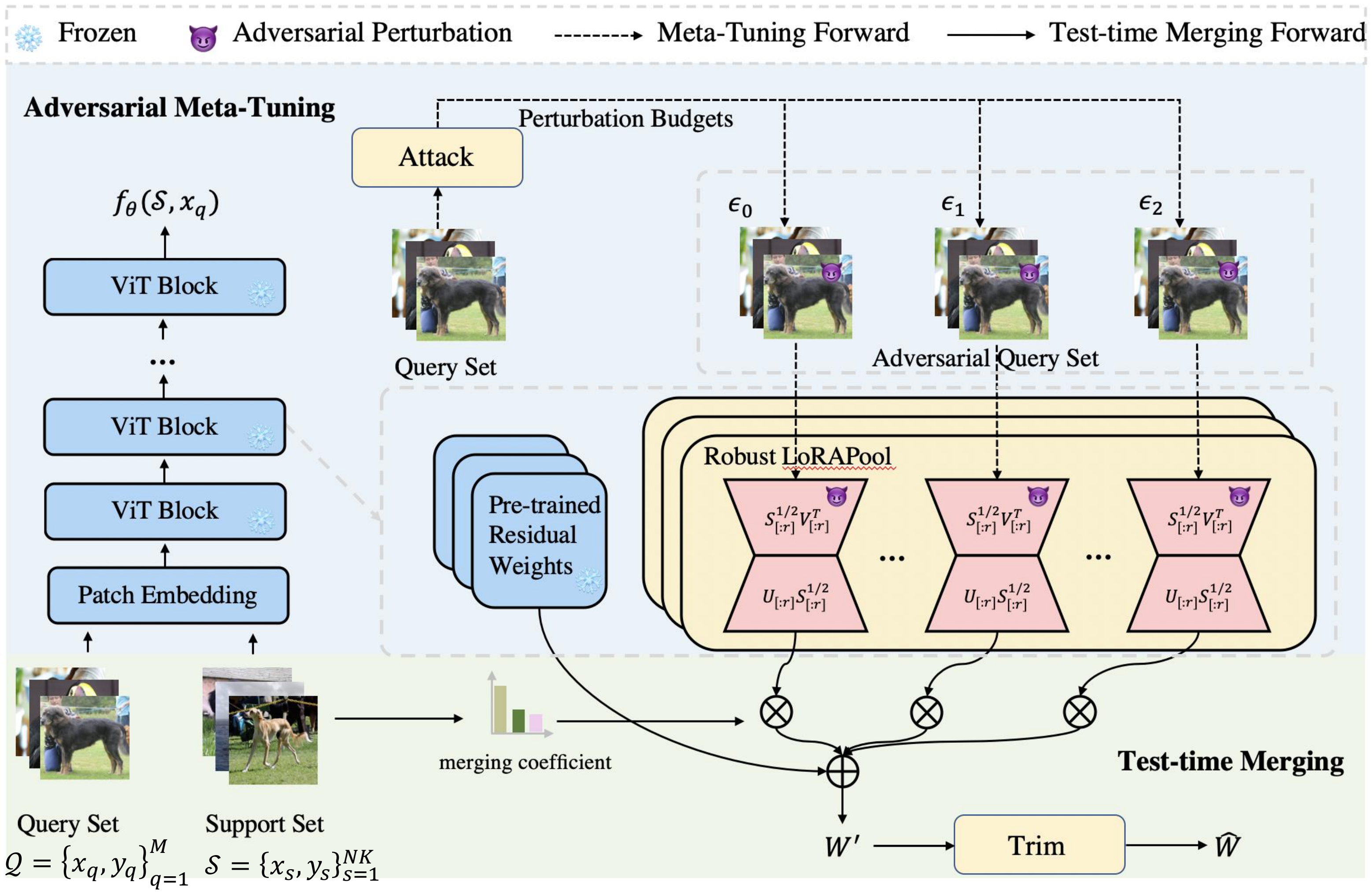
(Initialize LoRA parameters with the singular value decomposition results)

$$\delta_A = \eta_1 \cdot \frac{1}{M}\sum_{q=1}^M \nabla_A \mathcal{L}\left(f_{W^{\mathrm{res}}+AB}(\mathcal{S}, x_q^{adv}), y_q\right)$$

$$A \leftarrow A - \eta_2 \cdot \frac{1}{M}\sum_{q=1}^M \nabla_A \mathcal{L}\left(f_{W^{res}+(A+\delta_A)B}(\mathcal{S}, x_q^{adv}), y_q\right)$$

(Incorporate worst-case perturbation on A and B in the similar way)

Adaptive Robust LoRAPool



Weight Merging

support images *mean of class samples*

$$C_i = \frac{1}{NK}\sum_{s=1}^{NK}\gamma\left(\mathbf{f}_{W^{res}+A_iB_i}(x_s), \mathbf{p}_{i,y_s}\right)$$

$$V_i = \frac{1}{NK}\sum_{s=1}^{K}\sum_{\substack{c=1\\ c\neq y_s}}^{N}\gamma\left(\mathbf{f}_{W^{res}+A_iB_i}(x_s), \mathbf{p}_{i,c}\right)$$

cosine similarity

$$\zeta_i = \frac{\mathrm{Top}_k\left(\exp(-\beta(1-(\lambda C-(1-\lambda)V)))\right)_i}{\sum_{i=1}^{k}\mathrm{Top}_k\left(\exp(-\beta(1-(\lambda C-(1-\lambda)V)))\right)_i}$$

merging coefficient

$$W' = W^{res} + \sum_{i=1}^{P}\zeta_i A_i B_i.$$

Singular Value Trimming

$$\widehat{W} = \mathrm{trim}(W')$$

(Reset redundant singular values to zero)

Inference: Nearest-Centroid Classification

$$y_j^q = \underset{i}{\mathrm{argmin}}\, cos\left(\mathbf{f}_{\widehat{W}}\left(x_j^q\right), \mathbf{p}_{i,y_s}\right)$$

Clean ID/OOD Generalization Evaluation

Competitive methods: compared against three categories of related works:

- clean meta-tuning[1]
- parameter-efficient adaption[2]
- adversarial few-shot learning methods[3]

- **Does not sacrifice in-domain clean accuracy**
- **Good performance in clean OOD Generalization**

Few-shot Clean Accuracy on Meta-Dataset benchmark

1-shot	Backbone	TTF	In-domain	Out-of-domain									Avg.
			ImageNet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
PM [12]	ViT-small	-	65.07	59.03	38.13	76.18	61.56	57.29	56.03	80.41	55.17	54.42	60.35
StyleAdv [83]	ViT-small	-	56.10	62.25	40.38	66.62	55.94	57.93	53.19	81.10	54.20	48.08	57.58
AMT	ViT-small	-	**68.80**	**71.95**	**42.90**	**79.95**	**62.99**	**59.62**	**59.06**	**85.37**	**63.78**	**57.14**	**65.16**
PMF [12]	ViT-small	Y	65.07	71.52	38.67	76.15	61.62	59.82	56.03	80.41	59.71	54.41	62.34
PMF+AMT-FT	ViT-small	Y	**68.80**	**77.83**	**42.90**	**79.95**	**63.77**	**63.72**	**59.06**	**85.37**	**63.87**	**57.37**	**66.26**
ATTNSCALE [54]	ViT-small	Y	63.66	72.51	40.09	73.59	61.04	60.26	54.88	82.52	59.91	55.10	62.36
ATTNSCALE+AMT-FT	ViT-small	Y	**68.80**	**79.43**	**42.90**	**79.95**	**63.08**	**65.66**	**59.06**	**85.37**	**64.13**	**58.24**	**66.66**

5-shot	Backbone	TTF	In-domain	Out-of-domain									Avg.
			ImageNet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
PM [12]	ViT-small	-	80.71	78.77	56.56	92.23	79.92	76.16	76.98	96.61	74.66	71.77	78.44
StyleAdv [83]	ViT-small	-	74.51	80.22	58.78	87.60	78.67	75.57	73.80	96.18	71.99	63.93	76.12
AMT	ViT-small	-	**81.35**	**88.47**	**61.73**	**93.12**	**80.34**	**79.59**	**80.04**	**96.99**	**80.85**	**74.56**	**81.70**
PMF [12]	ViT-small	Y	79.92	93.54	67.45	92.22	**80.86**	81.64	77.25	96.61	87.68	75.33	83.25
PMF+AMT-FT	ViT-small	Y	**81.51**	**94.89**	**67.99**	**93.23**	80.41	**83.02**	**79.76**	**96.93**	**89.37**	**76.20**	**84.33**
ATTNSCALE [54]	ViT-small	Y	79.30	93.48	69.42	90.49	**81.04**	82.66	77.44	96.51	89.78	76.48	83.66
ATTNSCALE+AMT-FT	ViT-small	Y	**81.57**	**95.74**	**69.47**	**93.25**	80.96	**83.87**	**78.28**	**96.99**	**93.10**	**77.39**	**85.06**

Few-shot Clean Clean Accuracy on BSCD-FSL benchmark and fine-grained datasets

1-shot	Backbone	TTF	ChestX	ISIC	EuroSAT	CropDisease	CUB	Cars	Places	Plantae	Avg.
PM [12]	ViT-small	-	22.74 $\pm$0.40	33.72 $\pm$0.60	72.94 $\pm$0.77	81.04$\pm$0.85	83.53 $\pm$0.86	42.10 $\pm$0.80	71.66 $\pm$0.88	59.04$\pm$0.89	58.35
StyleAdv† [83]	ViT-small	-	**22.92$\pm$0.32**	33.05$\pm$0.44	72.15$\pm$0.65	81.22$\pm$0.61	84.01$\pm$0.58	40.48$\pm$0.57	72.64$\pm$0.67	55.52$\pm$0.66	57.75
AMT	ViT-small	-	22.39$\pm$0.39	**33.92 $\pm$0.58**	**73.52$\pm$0.84**	**82.04 $\pm$0.80**	**84.34 $\pm$0.83**	**44.33 $\pm$0.81**	**73.78$\pm$0.87**	**59.32 $\pm$0.94**	**59.21**
PMF [12]	ViT-small	Y	21.73$\pm$0.30	30.36$\pm$0.36	70.74$\pm$0.63	80.79$\pm$0.62	78.13$\pm$0.66	37.24$\pm$0.57	71.11$\pm$0.71	53.60$\pm$0.66	55.46
StyleAdv-FT† [83]	ViT-small	Y	22.92$\pm$0.32	**33.99$\pm$0.46**	**74.93$\pm$0.58**	**84.11$\pm$0.57**	84.01$\pm$0.58	40.48$\pm$0.57	72.64$\pm$0.67	55.52$\pm$0.66	58.57
AMT-FT	ViT-small	Y	**23.23$\pm$0.40**	33.95$\pm$0.63	73.95$\pm$0.78	82.04$\pm$0.8	**84.34$\pm$0.83**	**46.06 $\pm$0.80**	**73.83$\pm$0.89**	**59.32 $\pm$0.94**	**59.59**

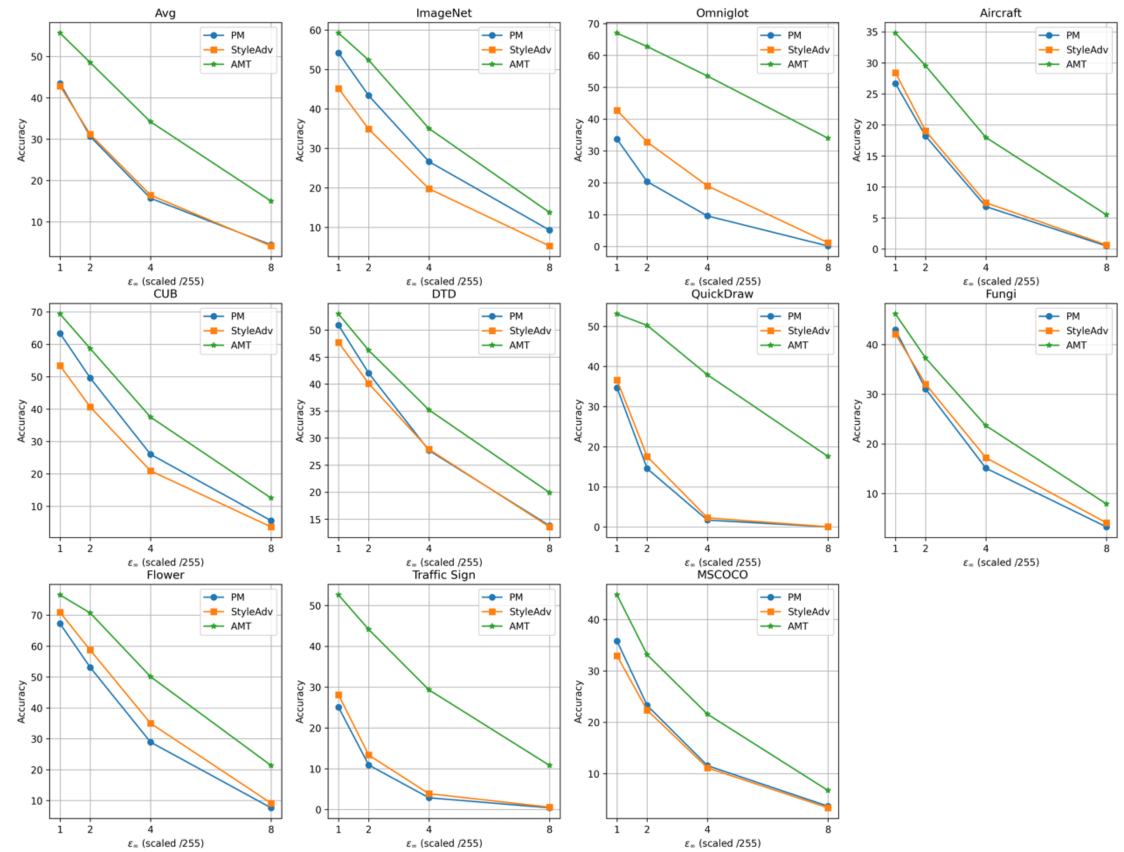
5-shot	Backbone	TTF	ChestX	ISIC	EuroSAT	CropDisease	CUB	Cars	Places	Plantae	Avg.
PM [12]	ViT-small	-	26.61 $\pm$0.43	47.60 $\pm$0.57	89.19$\pm$0.41	93.90$\pm$0.46	95.01 $\pm$0.40	**63.44$\pm$0.81**	88.73$\pm$0.51	78.31$\pm$0.71	72.85
StyleAdv† [83]	ViT-small	-	26.97$\pm$0.33	47.73$\pm$0.44	**88.57$\pm$0.34**	**94.85$\pm$0.31**	**95.82$\pm$0.27**	61.73$\pm$0.62	88.33$\pm$0.40	75.55$\pm$0.54	72.44
AMT	ViT-small	-	**27.54 $\pm$0.45**	**50.22$\pm$0.63**	88.38 $\pm$0.48	94.67 $\pm$0.40	94.86$\pm$0.39	62.94$\pm$0.82	**88.88$\pm$0.51**	**79.32$\pm$0.7**	**73.35**
PMF† [12]	ViT-small	Y	27.27	50.12	85.98	92.96	-	-	-	-	-
PMF [12]	ViT-small	Y	26.17$\pm$0.45	50.32 $\pm$0.63	89.97$\pm$0.40	94.77 $\pm$0.41	95.10$\pm$0.42	65.76$\pm$0.84	89.02$\pm$0.53	79.93$\pm$0.64	73.88
StyleAdv-FT† [83]	ViT-small	Y	26.97$\pm$0.33	51.23$\pm$0.51	90.12$\pm$0.33	**95.99$\pm$0.27**	**95.82$\pm$0.27**	66.02$\pm$0.64	88.33$\pm$0.40	78.01$\pm$0.54	74.06
AMT-FT	ViT-small	Y	**27.54 $\pm$0.45**	**51.56$\pm$0.68**	**90.62$\pm$0.40**	94.67 $\pm$0.40	95.21 $\pm$0.39	**67.18$\pm$0.79**	**89.22$\pm$0.50**	**80.36 $\pm$0.64**	**74.54**

[1] Pushing the Limits of Simple Pipelines for Few-Shot Learning: External Data and Fine-Tuning Make a Difference. In CVPR, 2022
[2] StyleAdv: Meta Style Adversarial Training for Cross-Domain Few-Shot Learning. In CVPR, 2023
[3] Strong Baselines for Parameter Efficient Few-Shot Fine-tuning. In AAAI, 2024

Robustness Against Double Distribution Shift

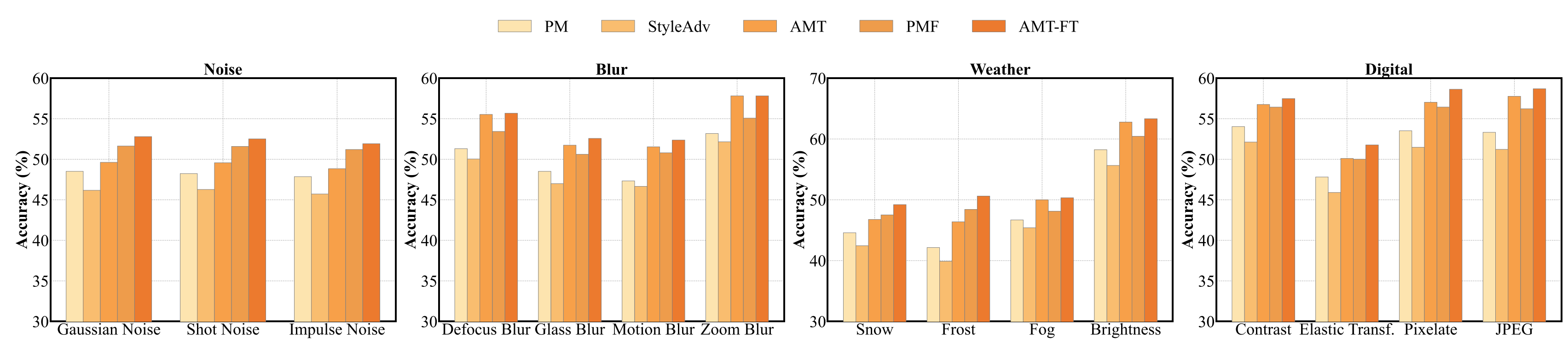
- **AMT handles adversarial attacks across varying perturbation budgets**

Few-shot Robustness Against Adversarial Attacks



- **AMT consistently outperforms counterparts across various common corruptions**

Few-shot Robustness Against Natural Corruptions



Average of *15 types of corruptions* × *5 multiple levels* × *10 domains*

Ablation Analysis

Component Effectiveness

- APQ: adversarial perturbation on query set
- APSV: adversarial perturbation on singular values and vectors
- RLP: Robust LoRAPool
- TTM: test-time merging
- STr: singular value trimming.

APQ	APSV	RLP	TTM	STr	In-domain	Out-of-domain									Avg.
					INet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
✗	✗	✗	✗	✗	65.07	59.03	38.13	76.18	61.56	57.29	56.03	80.41	55.17	54.42	60.35
✓	✗	✗	✗	✗	64.57	62.47	38.53	76.23	60.47	57.97	56.22	81.72	57.04	53.96	60.92
✓	✗	✓	✓	✗	65.56	63.92	39.74	76.06	61.73	58.64	55.99	80.93	56.96	54.28	61.38
✓	✗	✓	✓	✓	64.95	70.80	40.55	75.19	60.73	**59.66**	56.92	83.63	57.66	56.04	62.61
✓	✓	✓	✗	✗	67.95	62.16	39.13	79.27	61.77	58.75	56.59	79.74	55.45	54.63	61.54
✓	✓	✓	✓	✗	68.46	65.75	42.63	79.43	**63.10**	58.23	55.69	78.93	63.67	56.28	63.22
✓	✓	✓	✓	✓	**68.80**	**71.95**	**42.90**	**79.95**	62.99	59.62	**59.06**	**85.37**	**63.78**	**57.14**	**65.16**

Alternative Test-time Merging Strategies

Merging Strategies	In-domain	Out-of-domain									Avg
	INet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
Weight Average	63.96	64.49	40.57	74.23	59.99	57.36	55.73	80.50	59.87	53.04	60.97
Logit Average	65.84	**66.05**	40.70	78.72	60.79	**59.02**	**57.42**	**82.41**	58.41	55.09	62.44
Linear classifier	67.22	64.60	37.99	77.96	62.65	57.11	56.62	80.23	58.36	56.10	61.89
AMT	**68.46**	65.75	**42.63**	**79.43**	**63.10**	58.23	55.69	78.93	**63.67**	**56.28**	**63.22**

The Influence of Attack Pool Strategy

Method	In-domain	Out-of-domain									Avg.
	ImageNet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
Uniform LoRAPool	63.12	**73.28**	42.45	73.59	59.21	60.22	53.91	80.77	59.47	54.04	62.01
Random LoRAPool	64.30	72.28	**43.05**	79.03	58.75	**60.31**	57.15	84.02	60.01	**58.07**	63.70
Separate LoRAPool	**68.80**	71.95	42.90	**79.95**	**62.99**	59.62	**59.06**	**85.37**	**63.78**	57.14	**65.16**

Effective for Other Pre-training Methods

Method	In-domain	Out-of-domain									Avg.
	INet	Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	
DINO [1]	62.91	59.13	37.11	73.59	60.67	57.57	54.88	78.40	53.62	53.98	59.19
DINO+AMT	**68.80**	**71.95**	**42.90**	**79.95**	**62.99**	**59.62**	**59.06**	**85.37**	**63.78**	**57.14**	**65.16**
Δ	+5.89	+12.82	+5.79	+6.36	+2.32	+2.05	+4.18	+6.97	+10.16	+3.16	+5.97
iBOT [2]	65.09	61.57	35.40	70.85	60.36	57.37	54.47	78.04	55.00	55.00	59.32
iBOT+AMT	**69.95**	**69.89**	**38.84**	**79.96**	**61.65**	**62.35**	**58.34**	**79.67**	**61.88**	**56.49**	**63.90**
Δ	+4.86	+8.32	+3.44	+9.11	+1.29	+4.98	+3.87	+1.63	+6.88	+1.49	+4.58
DeIT [90]	74.23	57.32	35.20	69.36	51.73	56.08	45.52	64.31	53.82	54.64	56.22
DeIT+AMT	**81.11**	**65.50**	**38.36**	**75.80**	**56.53**	**62.16**	**53.19**	**76.09**	**58.98**	**58.57**	**62.63**
Δ	+6.88	+8.18	+3.16	+6.44	+4.80	+6.08	+7.67	+11.78	+5.16	+3.93	+6.41
AdvPre [25]	58.59	69.40	33.97	61.71	46.41	61.69	45.51	68.18	50.03	52.62	54.81
AdvPre+AMT	**73.35**	**73.72**	**37.16**	**69.79**	**52.41**	**63.87**	**49.91**	**75.62**	**59.69**	**56.16**	**61.17**
Δ	+14.76	+4.32	+3.19	+8.08	+6.00	+2.18	+4.40	+7.44	+9.66	+3.54	+6.36

Hyper-parameters Analysis

Loss Coefficient

(a) Clean Few-shot Accuracy

λ_{adv}	In-domain ImageNet	Out-of-domain Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	Avg.
0	68.50	70.95	41.53	79.74	62.02	59.29	**59.11**	84.72	56.14	56.57	63.85
6*	**68.80**	71.95	**42.90**	**79.95**	**62.99**	59.62	59.06	**85.37**	**63.78**	**57.14**	**65.16**
8	67.51	**72.23**	42.69	79.02	62.63	**59.97**	58.92	78.10	61.30	57.17	63.96

(b) Adversarial Few-shot Accuracy

λ_{adv}	In-domain ImageNet	Out-of-domain Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	Avg.
0	22.00	12.58	5.35	21.15	22.96	1.74	10.91	30.66	1.86	8.77	13.80
6*	**33.70**	42.19	11.72	32.05	32.47	27.45	19.74	41.12	22.79	17.67	28.09
8	31.85	**54.77**	**21.19**	**34.85**	**34.20**	**39.97**	**26.09**	**54.79**	**37.61**	**24.15**	**35.95**

LoRA Rank

r	In-domain ImageNet	Out-of-domain Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	Avg.
4	68.55	71.94	42.41	79.69	62.16	60.91	59.27	84.38	63.13	57.72	65.02
8*	**68.80**	71.95	42.90	79.95	62.99	59.62	59.06	**85.37**	**63.78**	57.14	**65.16**
16	68.22	72.15	**43.34**	79.98	62.43	60.86	56.64	83.67	62.97	57.13	64.74
32	68.29	71.96	43.00	81.11	63.07	**61.03**	**59.56**	80.50	63.29	**57.83**	64.96
64	67.39	72.20	43.15	81.21	62.98	60.56	56.74	83.54	62.90	57.13	64.78
128	68.35	**72.26**	42.74	**81.33**	**63.43**	60.62	56.70	83.86	63.25	57.09	64.96

Pool Size

P	ϵ mean	ϵ variance	In-domain ImageNet	Out-of-domain Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	Avg.
1	3.50	0	58.80	67.50	39.63	64.30	54.16	59.54	51.87	78.32	60.44	50.85	58.54
2	3.05	8.70	65.54	**72.62**	**43.39**	76.42	62.54	59.69	55.81	82.94	59.51	56.20	63.48
3	2.04	7.86	67.60	72.39	43.14	79.56	60.68	60.62	56.86	85.08	**63.88**	56.37	64.62
4*	3.53	12.56	**68.80**	71.95	42.90	79.95	62.99	59.62	**59.06**	**85.37**	63.78	57.14	**65.16**
5	3.52	10.05	67.18	71.26	42.76	**80.32**	**63.00**	**61.54**	58.53	82.56	61.71	**57.32**	64.62
6	4.02	11.85	65.73	71.48	42.53	73.99	60.87	59.84	55.46	85.18	60.67	55.93	63.17

Top-k

top-k	In-domain ImageNet	Out-of-domain Omglot	Acraft	CUB	DTD	QDraw	Fungi	Flower	Sign	COCO	Avg.
1	67.70	70.96	41.59	77.22	62.15	**61.05**	54.58	81.60	58.24	55.68	63.08
2*	**68.80**	**71.95**	42.90	79.95	**62.99**	59.62	**59.06**	**85.37**	**63.78**	57.14	**65.16**
3	68.29	73.11	**42.93**	**80.25**	62.73	60.56	58.03	82.94	61.61	**57.39**	64.78
4	65.97	71.89	42.65	78.50	61.80	60.12	57.43	84.84	61.83	57.38	64.24

Thanks!