Robust and Efficient Finetuning of Vision Foundation Models via Implicit Ensembling

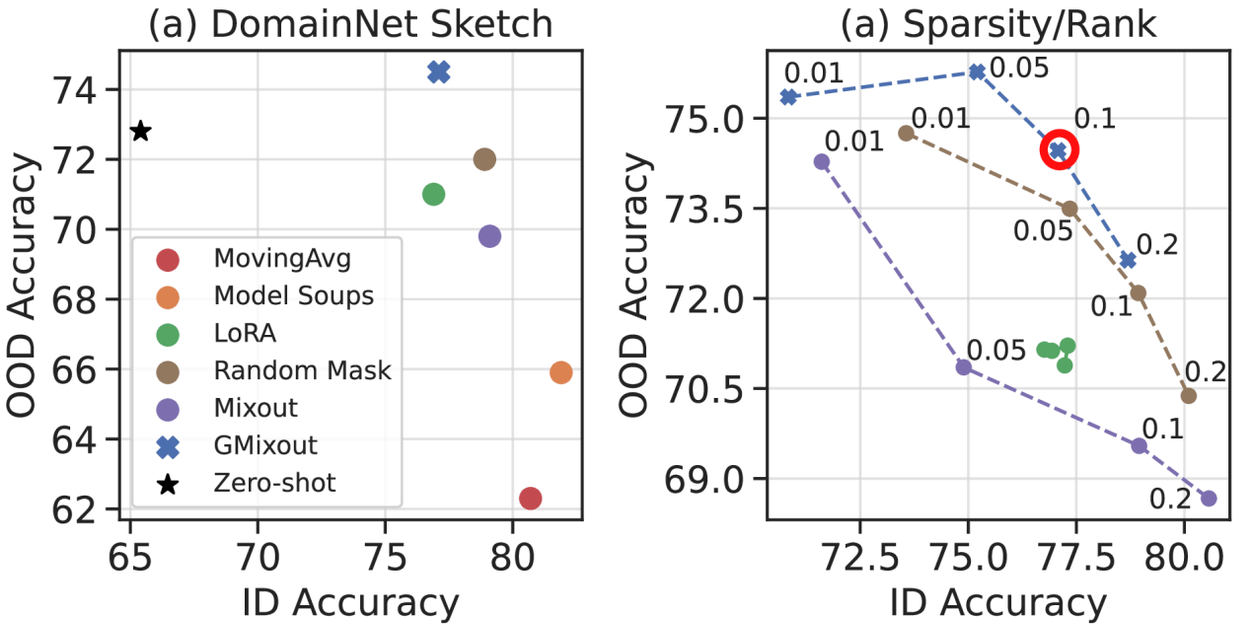
Masih Aminbeidokhti Heitor Rapela Medeiros Eric Granger Marco Pedersoli

LIVIA, ÉTS Montréal, Canada

International Laboratory on Learning Systems (ILLS)

Robust finetuning: adapt without losing robustness

- Finetune on ID data, stay robust to **unseen distribution shifts**.
- **Model Soups / ensembles**: robust, but many full runs.
- **PEFT** (LoRA, Random Mask): cheap, but limited OOD gains.
- **Mixout**: strong in NLP, **falls short for vision**.



ID vs. OOD accuracy on DomainNet (train on Sketch = ID; test on Real, Painting, Clipart = OOD). Left: one point per method (CLIP ViT-B/16), ★ = zero-shot CLIP; only GMixout ends up above zero-shot on OOD. Right: same, varying the fraction of trainable weights (labels; mask sparsity or LoRA rank). Upper-right is better.

Key idea: Mixout is an implicit ensemble

Mixout swaps finetuned weights with an anchor Φ_0 under a random mask $\mathbf{M} \sim \mathrm{Bernoulli}(1-p)$:

$$\min_{\Delta} \; \mathbb{E}_{\mathbf{M}} \left[\mathcal{L} \left(\tfrac{1}{1-p} \left((\mathbf{I} - \mathbf{M}) \odot \Phi_0 + \mathbf{M} \odot \Delta \right) \right) \right]$$

Each mask is a **subnetwork**; training averages over many of them. By bias–variance–covariance–locality:

$$\mathbb{E}\left[\mathcal{L}_{\text{ood}}(\Phi_{\text{WA}})\right] = \mathbb{E}\left[\text{bias}^2 + \tfrac{1}{M}\,\text{var} + \tfrac{M-1}{M}\,\text{cov}\right] + O(\bar{\Pi}^2)$$

- Mask resampled *every step* $\Rightarrow$ highly **correlated** subnetworks.
- Only lever is p: decorrelates, but also **pulls toward Φ_0**.
- **Fixed anchor** Φ_0 adds bias when the task is far from pretraining.

GMixout: generalized Mixout

1. **Resampling frequency k**: keep a mask for an *episode* of k steps $\Rightarrow$ diverse subnetworks. Mixout is $k = 1$.
2. **EMA anchor**: $\Phi_i = \lambda \Phi_{i-1} + (1 - \lambda)\Delta_i$ tracks a task-adapted solution.
3. **Sparse kernels**: update only ~10% of weights per step; merge at test time $\Rightarrow$ **no inference overhead**.

GMixout vs. **Mixout**

$\Delta \leftarrow \Phi_0$; $k \leftarrow 1$ or $k \leftarrow \lfloor T/E \rfloor$

for each iteration t **do**

  if $t \bmod k = 0$ **then**

    $\Phi_i \leftarrow \lambda \Phi_{i-1} + (1 - \lambda)\Delta$

    $\Delta \leftarrow \Phi_i$

  $\mathbf{M} \sim \text{Bernoulli}(1 - p)$

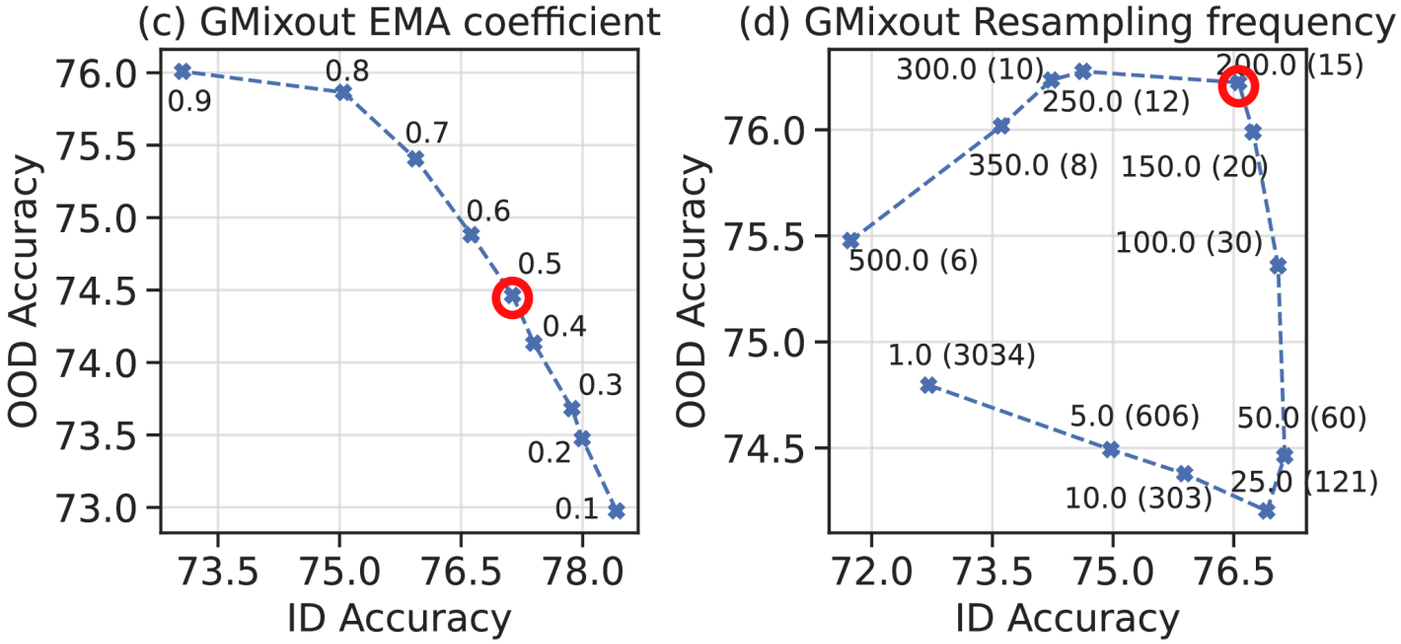
  update Δ with the Mixout objective

Experimental setup

Main experiments

- **Backbone**: CLIP ViT-B/16
- **Covariate shift**: DomainNet, iWildCam, ImageNet
- **Class imbalance**: ImageNet-LT, CIFAR100-LT
- **Baselines**: Full-FT, Model Soups, PEFT, Mixout, PACE, SoMA
- **GMixout**: $\lambda = 0.5$, k chosen on ID validation

Both knobs improve OOD accuracy



GMixout hyperparameter sweeps on DomainNet (train Sketch, test Real/Painting/Clipart). (c) EMA coefficient λ, point labels = λ. (d) Resampling interval k, labels = k (number of distinct masks / subnetworks seen in training). Upper-right is better.

- EMA coefficient λ: OOD accuracy from **73.0 to 76.0** on DomainNet-Sketch.
- Resampling interval k: OOD from 74.8 ($k = 1$) to **76.2** while staying competitive on ID.

Covariate shift: DomainNet & iWildCam

Method	DomainNet (OOD acc.)					iWildCam (F1)	
	Real	Sketch	Paint.	Clip.	Avg.	ID	OOD
Zero-shot	65.8	72.8	71.4	70.4	70.1	11.5	12.7
Full-FT	60.2	62.4	61.6	61.9	61.5	46.5	36.4
Model Soups	64.3	65.9	64.6	65.2	65.0	46.8	37.6
LoRA	66.5	71.0	69.0	69.5	69.0	49.7	35.9
Mixout	66.5	69.8	65.0	67.7	67.3	43.5	32.9
PACE	67.5	74.0	71.6	72.9	71.5	43.1	32.3
SoMA	67.3	73.8	70.8	72.9	71.2	50.8	36.2
GMixout	**68.3**	**75.3**	**73.5**	**73.4**	**72.6**	**51.8**	**38.4**

+5.3 over Mixout and **+2.6** over zero-shot on DomainNet; best ID *and* OOD on iWildCam.

ImageNet, class imbalance, and cost

ImageNet (OOD avg. over V2/R/S/A)

Method	ID	OOD
Zero-shot	68.2	59.1
Model Soups	**84.4**	60.0
LoRA	83.7	59.9
Mixout	82.7	54.4
GMixout	82.8	**60.7**

Long-tailed (balanced acc.)

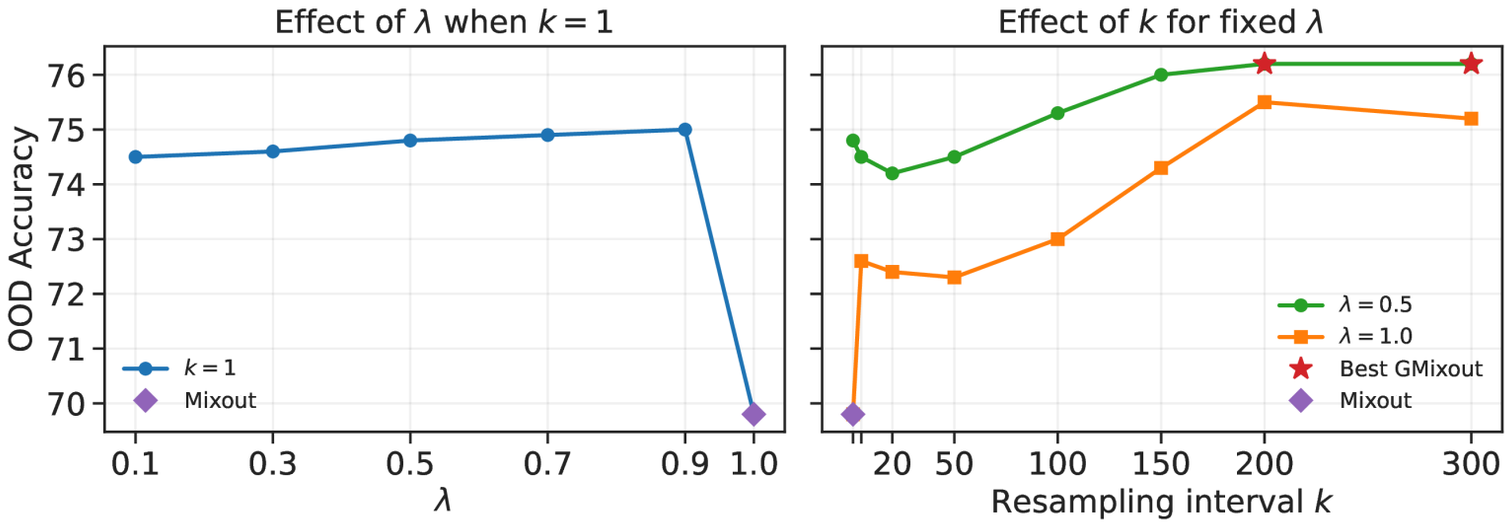
Method	IN-LT	C100-LT
Random Mask	76.8	81.5
Mixout	70.9	81.5
GMixout	**77.0**	**82.2**

Per-step training cost (ViT-B/16)

Method	Params (M)	FLOPs (T)	VRAM (G)
Full-FT	85.5	1.7	3.2
Model Soups	427.5	8.5	16.0
LoRA	8.7	1.3	2.3
GMixout	9.0	1.1	2.6

- ViT-L/14@336 finetuned on a **single 24GB RTX 3090**.
- Same inference cost as a single dense model.

Why does it work?



One component at a time, DomainNet-Sketch (OOD accuracy on Real/Painting/Clipart). Left: EMA anchor only, vary λ with the mask resampled every step ($k = 1$). Right: vary resampling interval k with EMA ($\lambda = 0.5$) or without it ($\lambda = 1$). ◆ = standard Mixout ($k = 1, \lambda = 1$); ★ = best GMixout (76.2).

- EMA anchor alone and longer resampling alone each beat Mixout; **combined: 76.2**.
- Episodic EMA *without* masking reaches only 68.3 $\Rightarrow$ masked subnetworks matter.

Takeaways

- Mixout is a **single-run implicit ensemble**; anchor, resampling frequency and sparsity control the ID–OOD trade-off.
- **GMixout** adds an EMA anchor, a resampling frequency and sparse kernels.
- Beats Mixout, PEFT and Model Soups across 5 benchmarks, **no inference overhead**.



github.com/Masseeh/GMixout