Enhancing Cross-Modal Fine-Tuning with Gradually Intermediate Modality Generation

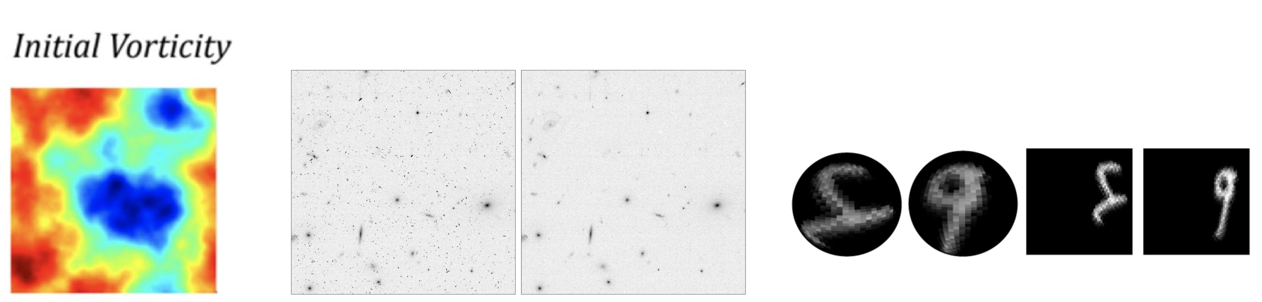
Lincan Cai[1] Shuang Li[1,✉] Wenxuan Ma[1] Jingxuan Kang[2] Binhui Xie[1] Zixun Sun[3] Chengwei Zhu[3]

[1] Beijing Institute of Technology [2] University of Illinois Urbana Champaign [3] Tencent

Email: lincancai@bit.edu.cn & shuangli@bit.edu.cn

Motivation & Contribution

- In certain professional field where data is **scarce and modality specific**, the need for specialized models becomes apparent.

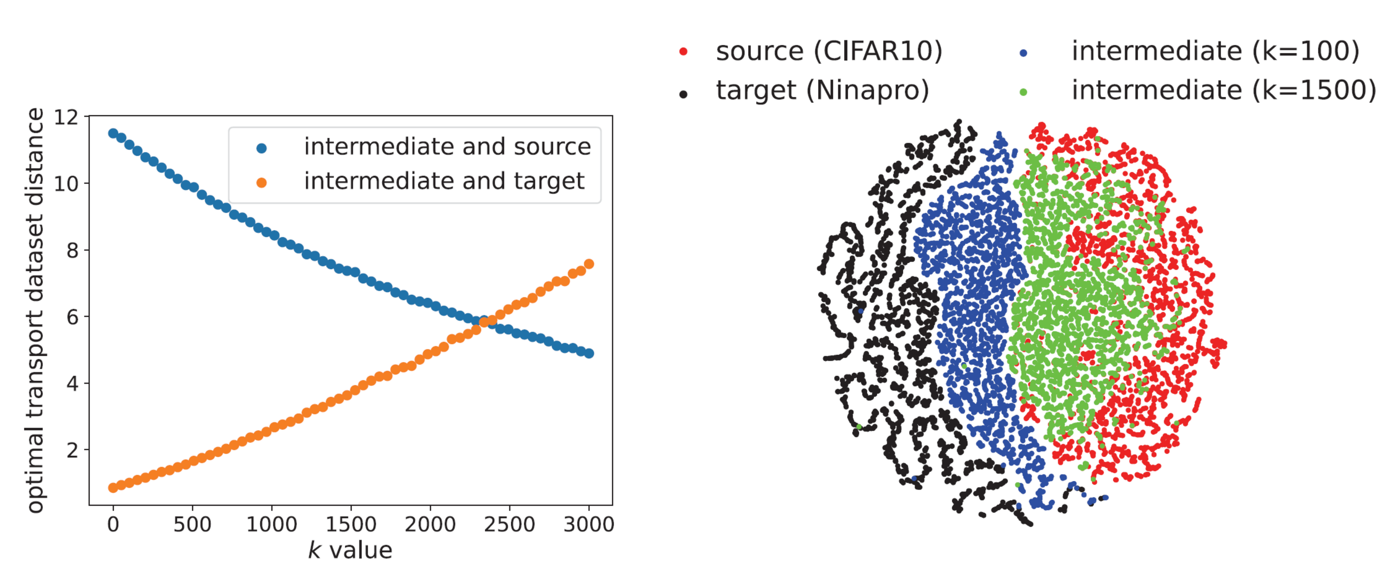


- Directly pre-training a large model **from scratch** on the target modality **performs poorly**.

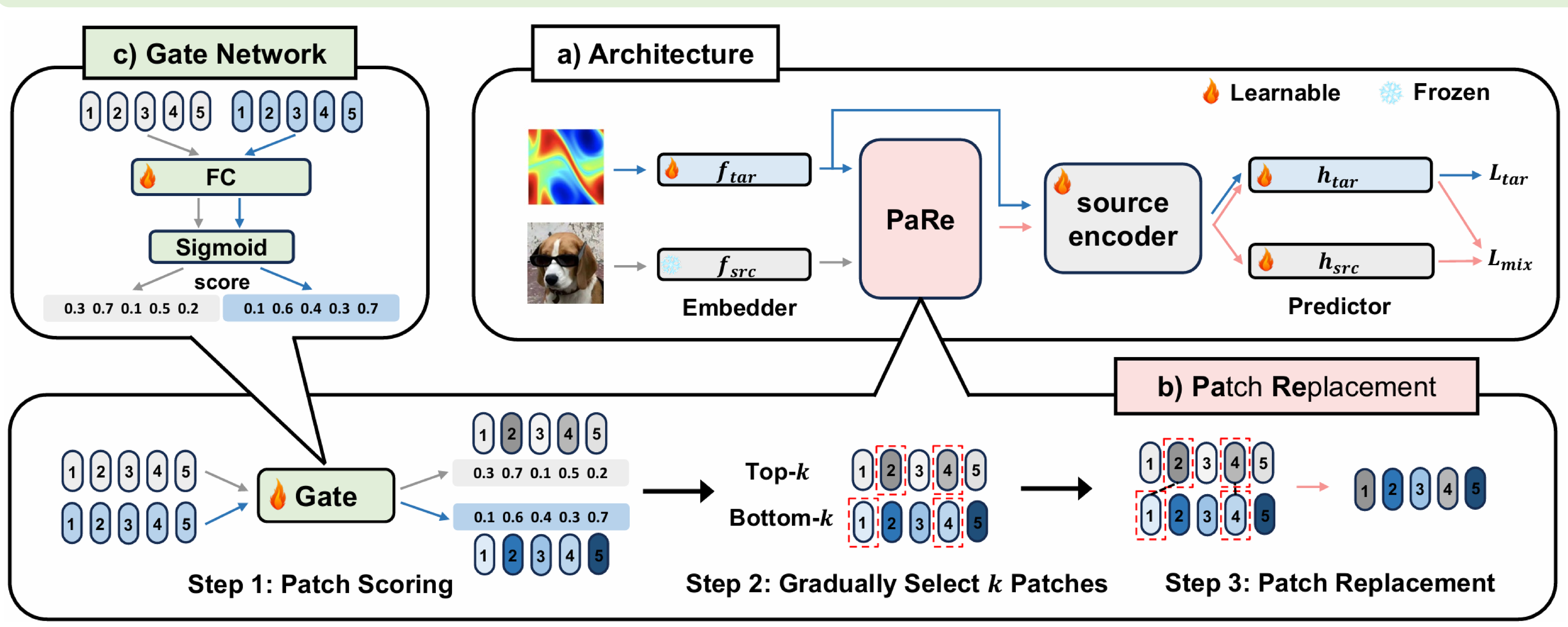
Error_Rate	NinaPro	PSICOV	Cosmic
Type	Gesture recognition	Protein prediction	Cosmic rays
Data Size	3956	3606	5250
Pretrained model	9.96	5.09	0.5
Specialized model	6.60	2.94	0.127

- There is a **significant disparity** between the source and target modalities, and their **label spaces** do not overlap at all, it is difficult to leverage pretrained large models on image or text data with rich annotations for **cross-modal transfer**.

- We gradually constructs **intermediate modalities** from the source modality to the target modality, **bridging the modality gap**.
- By **mixing** the source modality data with the target modality data to construct intermediate modality data, we can also alleviates the issue of **insufficient data** volume in the target modality.
- Utilize **Curriculum Learning**, allowing the model to transition from intermediate modality data that is closer to the source modality to that is closer to the target modality. This enables a gradual transfer from **easy to difficult** tasks.



Framework



✓ **Architecture Design**
- Source and target modality embedder f^s and f^t.
- Source and target modality predictor h^s and h^t.
- Pretrained feature encoder g.

✓ **Gate Network**
- A **Fully Connected** Layer followed by a **Sigmoid** Layer.
- Ensuring gradient flow during backpropagation using **Gumble Softmax** function.
- Effectively selects the **most informative** patches of both source and target modality for classification.

✓ **Traditional Mixing Methods** ☹
- **Mixup:** when applied across modalities with significant differences, it can confuse the model due to the disparate nature of the modalities.
- **Cutmix:** different modalities often have different important regions, making it difficult to identify and preserve the critical features from each modality.

✓ **Loss Function**

$$\mathcal{L}_{mix} = (1-\lambda)\mathcal{L}_{tar}(p^{mt}, y^t) + \lambda\mathcal{L}_{src}(p^{ms}, y^s)$$

$$\mathcal{L}_{total} = \beta_1\mathcal{L}_{tar}(p_t, y_t) + \beta_2\mathcal{L}_{mix}$$

✓ **Modality-Agnostic Patch Replacement**
- **Patch Scoring:** use the gate network to score each patch from source and target modality.
- **Patch Selection:** Select the Top-k highest-scoring patches from source modality and Bottom-k lowest-scoring patches from the target modality.
- **Patch Replacement:** Replace the selected bottom-k target modality patches with the selected top-k source modality patches to construct intermediate modality data.

Paper Code

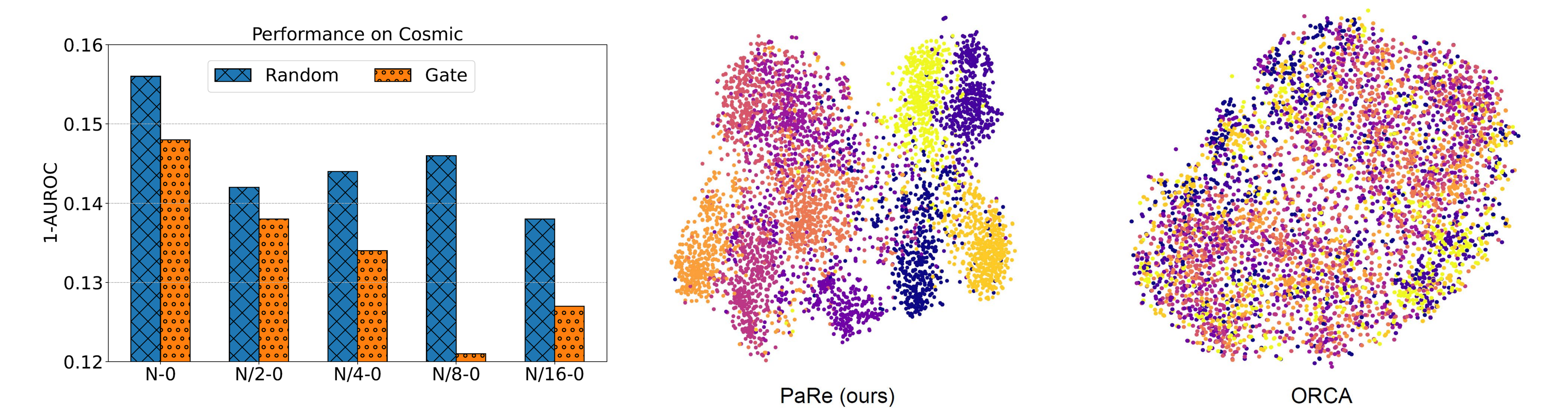
Experiment

Table 1: Prediction errors (↓) across 10 diverse tasks on NAS-Bench-360. "FPT" and "NFT" respectively represent fine-tuning only the layer normalization of the model and performing one-stage full fine-tuning of the model.

	CIFAR-100 0-1 error (%)	Spherical 0-1 error (%)	Darcy Flow relative ℓ_2	PSICOV MAE_8	Cosmic 1-AUROC	NinaPro 0-1 error (%)	FSD50K 1-mAP	ECG 1-F1 score	Satellite 0-1 error (%)	DeepSEA 1-AUROC
Hand-designed	19.39	67.41	8.00E-03	3.35	0.127	8.73	0.62	0.28	19.80	0.30
NAS-Bench-360	23.39	48.23	2.60E-03	2.94	0.229	7.34	0.60	0.34	12.51	0.32
DASH	24.37	71.28	7.90E-03	3.30	0.190	6.60	0.60	0.32	12.28	0.28
Perceiver IO	70.04	82.57	2.40E-02	8.06	0.485	22.22	0.72	0.66	15.93	0.38
FPT	10.11	76.38	2.10E-02	4.66	0.233	15.69	0.67	0.50	20.83	0.37
NFT	7.67	55.26	7.34E-03	1.92	0.170	8.35	0.63	0.44	13.86	0.51
ORCA	6.53	29.85	7.28E-03	1.91	0.152	7.54	0.56	**0.28**	11.59	0.29
PaRe	**6.25**	**25.55**	**7.00E-03**	**0.99**	**0.121**	**6.53**	**0.55**	**0.28**	**11.18**	**0.28**

Table 2: Normalized Root Mean Squared Errors (nRMSEs, ↓) across 8 tasks of PDEBench. PaRe surpasses U-Net and PINN in all tasks, outperforms ORCA in 6 out of 8 tasks, and exhibits performance comparable to FNO.

	Advection 1D	Burgers 1D	Diffusion-Reaction 1D	Diffusion-Sorption 1D	Navier-Stokes 1D	Darcy-Flow 2D	Shallow-Water 2D	Diffusion-Reaction 2D
PINN	6.70E-01	3.60E-01	6.00E-03	1.50E-01	7.20E-01	1.80E-01	8.30E-02	8.40E-01
FNO	1.10E-02	**3.10E-03**	**1.40E-03**	1.70E-03	6.80E-02	2.20E-01	**4.40E-03**	**1.20E-01**
U-Net	1.10E+00	9.90E-01	8.00E-02	2.20E-01	-	-	1.70E-02	1.60E+00
ORCA	9.80E-03	1.20E-02	3.00E-03	**1.60E-03**	**6.20E-02**	8.10E-02	6.00E-03	8.20E-01
PaRe	**2.70E-03**	8.30E-03	2.60E-03	**1.60E-03**	6.62E-02	**8.06E-02**	5.70E-03	8.18E-01



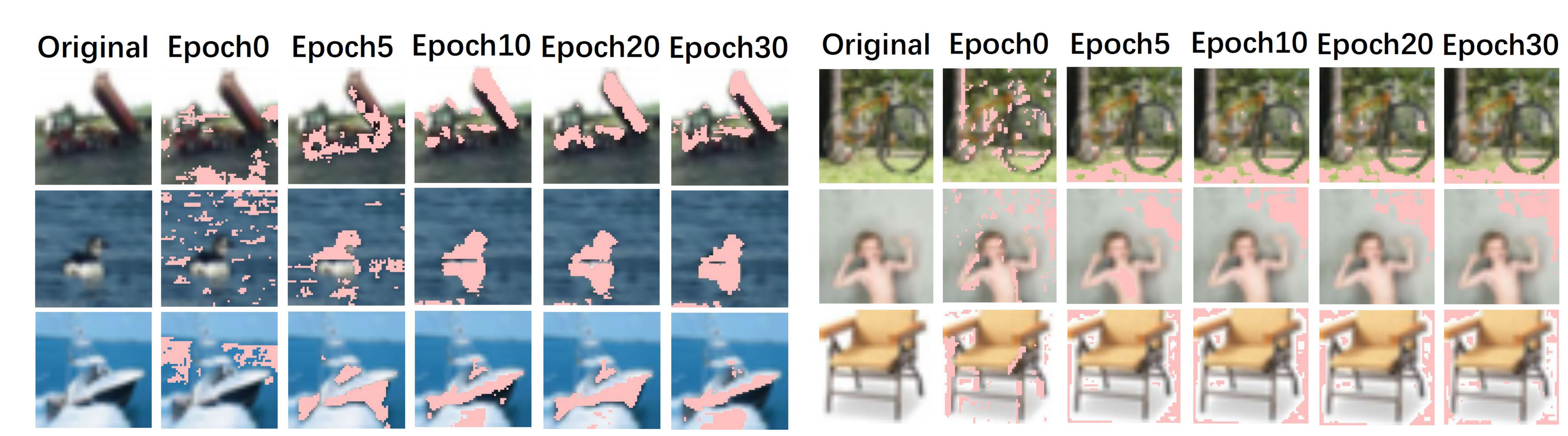


Table 4: Comparison prediction errors (↓) of traditional mixing strategies and PaRe variants across 10 diverse tasks, and the impact of varying strategies for different values of k, where "non-gradual" indicates a constant k, while the other three represent different strategies for decreasing k.

	Method	CIFAR-100	Spherical	Darcy Flow	PSICOV	Cosmic	NinaPro	FSD50K	ECG	Satellite	DeepSEA
	Mixup	6.59	26.60	7.70E-03	**0.99**	0.500	7.74	0.56	0.29	11.51	0.29
	CutMix	**6.11**	27.76	7.20E-03	**0.99**	0.135	8.41	0.56	0.29	11.58	**0.28**
PaRe	w/ non-gradual	6.59	27.68	7.20E-03	**0.99**	0.138	7.59	0.57	0.28	11.61	0.29
	w/ piecewise	6.22	26.88	**6.90E-03**	**0.99**	0.132	6.98	0.56	0.29	**10.89**	0.29
	w/ exponential	6.38	26.35	7.00E-03	**0.99**	**0.119**	7.13	**0.55**	**0.28**	11.56	**0.28**
	w/ linear (default)	6.25	**25.55**	7.00E-03	**0.99**	0.121	**6.53**	**0.55**	**0.28**	11.18	**0.28**

Table 5: Comparison prediction errors (↓) between different strategies (Random vs. Gate) to select patches for replacement.

	CIFAR-100	Spherical	Darcy Flow	PSICOV	Cosmic	NinaPro	FSD50K	ECG	Satellite	DeepSEA
Random	6.52	28.06	7.10E-03	**0.99**	0.146	6.98	0.56	0.29	11.32	**0.28**
Gate	**6.25**	**25.55**	**7.00E-03**	**0.99**	**0.121**	**6.53**	**0.55**	**0.28**	**11.18**	**0.28**