Residual Feature Integration is Sufficient to Prevent Negative Transfer

Yichen Xu[1*] Ryumei Nakada[2*] Linjun Zhang[3†] Lexin Li[1]

[1]University of California, Berkeley
[2]Harvard University
[3]Rutgers University

{yichen_xu, lexinli}@berkeley.edu
lz412@stat.rutgers.edu
ryumei_nakada@hms.harvard.edu

Main message: **a frozen transferred representation + a trainable residual encoder** is enough to **avoid negative transfer in rate** and adapt to transfer quality.

Motivation: transfer can hurt

Problem. Standard transfer uses a frozen source representation

$$x \mapsto f_{\mathrm{rep}}(x), \qquad g(x) = w^\top f_{\mathrm{rep}}(x),$$

or a lightweight adapter on top of $f_{\mathrm{rep}}(x)$.

Failure mode: negative transfer

$$\mathcal{R}_{\mathbb{P}^t}(\widehat{g}_{\mathsf{transfer}}) > \mathcal{R}_{\mathbb{P}^t}(\widehat{g}_{\mathsf{scratch}}).$$

This happens under:

- distribution shift,
- task mismatch,
- corrupted source supervision,
- missing modality at adaptation time.

Question

Can we use transferred features when they help, but fall back to scratch learning when they do not?

Method: Residual Feature Integration (REFINE)

$$g(x) = v^\top f_{\text{rep}}(x) + u\, h(x)$$

- $f_{\text{rep}}(x)$: frozen source-side representation
- $h(x)$: trainable target-side encoder
- $v^\top f_{\text{rep}}(x)$: linear probe on transferred features
- $u\, h(x)$: residual correction learned from target data

Train only h and the shallow head; keep the source model fixed.

Interpretation

- If transfer is good: h learns a small residual.
- If transfer is bad: h recovers target-only learning.

Theoretical setup

We work in nonparametric regression:

$$Y_i = f^*(X_i) + \varepsilon_i, \qquad X_i \in [0,1]^d, \qquad f^* \in \mathcal{C}^\beta.$$

Define the best linear probe on the transferred representation:

$$v^* = \arg\min_{v \in \mathbb{R}^p} \mathbb{E}\big(v^\top f_{\mathrm{rep}}(X) - f^*(X)\big)^2.$$

Residual difficulty is measured by

$$\rho^* := \left\| v^{*\top} f_{\mathrm{rep}} - f^* \right\|_{\mathcal{C}^\beta}.$$

Key quantity

ρ^* is the β-Hölder norm of the residual after the best linear probe. Small ρ^*: informative transfer. Large ρ^*: mismatch.

Model class and ERM

Residual encoder h is a clipped ReLU network. The REFINEclass is

$$\mathcal{G}_{d,\rho}(W, L, B; f_{\mathrm{rep}}) = \Big\{ g(x) = v^\top f_{\mathrm{rep}}(x) + u\, h(x) : |u| \le 1,\ \|v\| \le 1,\ h \in \bar{\mathcal{H}}_d(W, L, B) \Big\}.$$

Train by empirical risk minimization:

$$\widehat{g} = \arg \min_{g \in \mathcal{G}_{d,\rho}(W,L,B;f_{\mathrm{rep}})} \frac{1}{n} \sum_{i=1}^{n} (g(X_i) - Y_i)^2.$$

Capacity choice in the theorem:

$$L = c_1, \qquad W = c_2 \max\{n^{d/(2\beta+d)} \rho^{2d/(2\beta+d)}, 1\},$$

$$B = (\rho^* \vee 1) \max\{n\rho^2, 1\}^{c_3}.$$

Here ρ is a tuning parameter controlling residual-network complexity.

Main theorem: excess risk bound

Assume $v^{*\top} f_{\mathrm{rep}} - f^* \in \mathcal{C}^\beta$. Then

$$\mathbb{E}[\mathcal{R}_{\mathbb{P}^t}(\widehat{g}) - \mathcal{R}_{\mathbb{P}^t}(f^*)] \le C\left\{ \left(\rho^{\frac{2d}{2\beta+d}} \log n + \rho^{*2}\rho^{-\frac{4\beta}{2\beta+d}}\right) n^{-\frac{2\beta}{2\beta+d}} + \frac{p \log n}{n} \right\}.$$

Structure of the bound

- $\frac{p \log n}{n}$: parametric cost of estimating the linear probe v
- $n^{-2\beta/(2\beta+d)}$: nonparametric rate for learning the residual
- ρ: bias–variance tradeoff for the residual network
- ρ^*: quality of the transferred representation

Corollary 1: fallback to scratch rate

For any fixed $\rho > 0$,

$$\mathbb{E}[\mathcal{R}_{\mathbb{P}^t}(\widehat{g}) - \mathcal{R}_{\mathbb{P}^t}(f^*)] = \widetilde{O}\Big(n^{-\frac{2\beta}{2\beta+d}} + \frac{p}{n}\Big).$$

Meaning

ReFineis never worse, up to log factors, than the standard minimax rate for learning a β-Hölder target from scratch.

So if transfer is uninformative, the residual path recovers the target-only regime.

Corollary 2: adaptivity to representation quality

Balancing the theorem by choosing $\rho = \rho^*$ yields

$$\mathbb{E}[\mathcal{R}_{\mathbb{P}^t}(\widehat{g}) - \mathcal{R}_{\mathbb{P}^t}(f^*)] = \widetilde{O}\left(\rho^{* \frac{2d}{2\beta+d}} n^{-\frac{2\beta}{2\beta+d}} + \frac{p}{n}\right).$$

Meaning

- Small ρ^*: the residual task is easy, so the nonparametric term shrinks.
- Large ρ^*: the rate falls back to the classical nonparametric rate.

This is the precise sense in which the rate transitions from nonparametric to near-parametric.

Corollary 3: formal no-negative-transfer guarantee

Define

$$\mathcal{F}^{\beta}(f_{\mathrm{rep}}, \gamma) = \left\{ f^* : [0,1]^d \to \mathbb{R} : \min_{\|v\| \le 1} \|v^\top f_{\mathrm{rep}} - f^*\|_{\mathcal{C}^\beta} \le \gamma \right\}.$$

Then, for scratch ERM $\widehat{g}_{\mathrm{sc}}$ and linear probe $\widehat{w}_{\mathrm{ft}}$,

$$\sup_{f^* \in \mathcal{F}^{\beta}(f_{\mathrm{rep}}, \gamma)} \mathbb{E}[\mathcal{R}_{\mathbb{P}^{\mathrm{t}}}(\widehat{g}) - \mathcal{R}_{\mathbb{P}^{\mathrm{t}}}(f^*)] = \widetilde{O}\left(\min \left\{ \sup_{f^* \in \mathcal{F}^{\beta}(f_{\mathrm{rep}}, \gamma)} \mathbb{E}[\mathcal{R}_{\mathbb{P}^{\mathrm{t}}}(\widehat{g}_{\mathrm{sc}}) - \mathcal{R}_{\mathbb{P}^{\mathrm{t}}}(f^*)], \right. \right.$$

$$\left. \left. \sup_{f^* \in \mathcal{F}^{\beta}(f_{\mathrm{rep}}, \gamma)} \mathbb{E}\left[\mathcal{R}_{\mathbb{P}^{\mathrm{t}}}(\widehat{w}_{\mathrm{ft}}^\top f_{\mathrm{rep}}) - \mathcal{R}_{\mathbb{P}^{\mathrm{t}}}(f^*)\right] \right\} \right).$$

Meaning

Over the target-function class induced by f_{rep}, REFINE matches the better of scratch learning and linear probing, up to log factors.

Proof sketch

Start from the decomposition

$$f^*(x) = \underbrace{v^{*\top} f_{\mathrm{rep}}(x)}_{\text{linear transferred part}} + \underbrace{\left(f^*(x) - v^{*\top} f_{\mathrm{rep}}(x)\right)}_{\text{residual } r^*(x)}.$$

Step 1. Estimate v^* by a linear probe:

$$\text{error} \sim \frac{p}{n}.$$

Step 2. Approximate and estimate $r^* \in \mathcal{C}^\beta$ by a ReLU network:

$$\text{error} \sim \left(\rho^{\frac{2d}{2\beta+d}} \log n + \rho^{*2} \rho^{-\frac{4\beta}{2\beta+d}}\right) n^{-\frac{2\beta}{2\beta+d}}.$$

Step 3. Combine the two errors.

If f_{rep} is useless, then $r^* \approx f^*$: recover scratch learning. If f_{rep} is good, then r^* is small: only a small correction is needed.

Experiments settings

Benchmarks

- Vision: CIFAR-10/100, STL, Clipart, Sketch, USPS, MNIST
- Text: Books, DVD, Electronics, Kitchen
- Tabular: additional appendix experiments

Baselines

- NoTrans
- LinearProbe
- Adapter
- Distillation
- LoRA
- DANN-Gate

Stress tests

- 40% and 80% label flips
- semantic perturbation
- class imbalance

Experiments natural shift results

Dataset	Method	Acc	AUC	F1	MinCAcc
CIFAR100→10	NoTrans	**56.58**	**0.9005**	**0.5634**	**37.20**
	LinearProb	38.93	0.8284	0.3815	16.94
	Adapter	38.23	0.8247	0.3754	16.46
	LoRA	43.14	0.8603	0.4237	20.14
	DANN-Gate	43.22	0.8605	0.4214	17.48
	REFINE	54.40	0.8942	0.5406	33.62
CIFAR10→100	NoTrans	18.32	0.8140	0.1774	1.00
	LinearProbe	7.01	0.7489	0.0496	0.00
	Adapter	6.56	0.7499	0.0459	0.00
	LoRA	6.82	0.7558	0.0463	0.00
	DANN-Gate	5.20	0.7341	0.0285	0.00
	REFINE	**18.59**	**0.8276**	**0.1787**	**1.40**

Experiments natural shift results

Dataset	Method	Acc	AUC	F1	MinCAcc
CIFAR10→STL	NoTrans	48.69	0.8683	0.4831	**26.80**
	LinearProbe	50.27	0.8795	0.4955	18.93
	Adapter	49.29	0.8773	0.4865	15.68
	LoRA	50.76	0.8813	0.4930	5.68
	DANN-Gate	47.71	0.8659	0.4712	13.93
	ReFine	**53.42**	**0.8944**	**0.5301**	25.98
Clipart→Sketch	NoTrans	18.88	0.7170	0.1828	0.00
	LinearProbe	18.34	0.7290	0.1727	0.00
	Adapter	18.24	**0.7369**	0.1549	0.00
	LoRA	16.90	0.6937	0.1671	0.00
	DANN-Gate	16.58	0.6942	0.1544	0.00
	ReFine	**20.34**	0.7338	**0.1968**	**0.53**

Experiments natural shift results

Dataset	Method	Acc	AUC	F1	MinCAcc
USPS→MNIST	NoTrans	62.07	0.9566	0.5967	9.29
	LinearProbe	67.00	0.9469	0.6563	9.16
	Adapter	61.87	0.9375	0.5952	8.88
	LoRA	64.82	0.9333	0.6435	29.33
	DANN-Gate	52.21	0.9012	0.4853	0.02
	REFINE	**70.05**	**0.9582**	**0.6954**	**31.62**
Books→Kitchen	NoTrans	71.66	0.7848	0.7161	**68.60**
	LinearProbe	66.74	0.7568	0.6571	51.56
	Adapter	71.34	0.7839	0.7111	62.88
	LoRA	66.96	0.7279	0.6695	65.64
	DANN-Gate	66.60	0.7330	0.6659	64.68
	REFINE	**72.72**	**0.8147**	**0.7248**	65.52

Experiments natural shift results

Dataset	Method	Acc	AUC	F1	MinCAcc
DVD→Electronics	NoTrans	68.52	0.7585	0.6806	59.80
	LinearProbe	66.06	0.7266	0.6580	58.36
	Adapter	65.86	0.7206	0.6577	61.44
	LoRA	66.56	0.7170	0.6656	**65.40**
	DANN-Gate	66.90	0.7196	0.6686	63.56
	REFINE	**70.34**	**0.7886**	**0.6995**	61.72

Adapt-time multimodality extension

A distinct transfer setting: a new modality appears only at adaptation time.

- Spatial transcriptomics: each cell has transcriptome + spatial coordinates
- Task: lymph-node anatomical domain classification
- Source model: scGPT pretrained on dissociated RNA only
- So spatial information is *absent during pretraining* but available at adaptation time

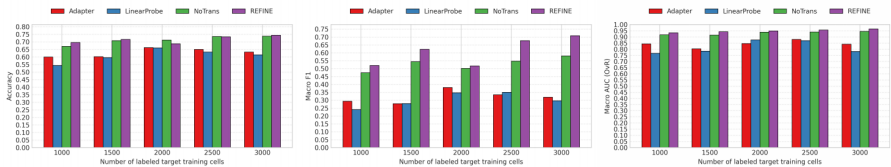
This creates an **adapt-time multimodality extension** problem: the model must incorporate a modality that the pretrained representation never observed.

$$g(x) = v^\top f_{\mathrm{rep}}(x_{\mathrm{RNA}}) + u\, h(x_{\mathrm{RNA}}, x_{\mathrm{spatial}})$$

Key idea

REFINE keeps the frozen scGPT representation and uses the residual branch to inject the missing spatial modality at adaptation time.

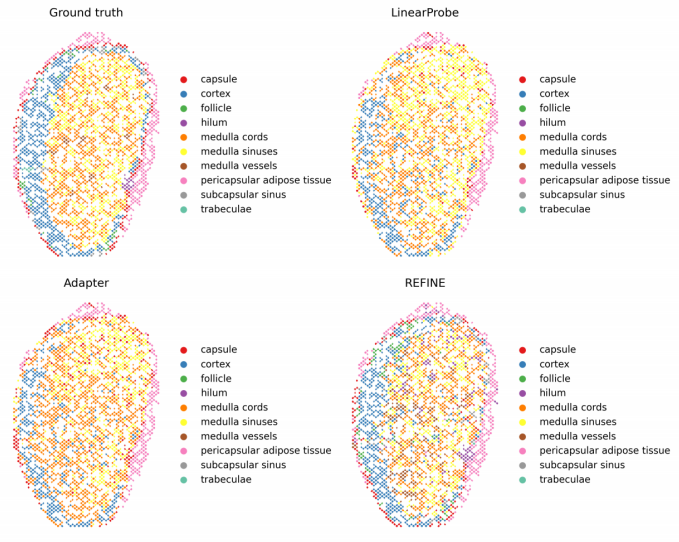
Adapt-time multimodality extension: Fig. 2



ACC F1 AUC

Figure 2: Metric comparison across labeled target sizes for Adapter, LinearProbe, NoTrans, and REFINE.

Adapt-time multimodality extension: Fig. S2



Takeaways

1. **Method:**
$$g(x) = v^\top f_{\mathrm{rep}}(x) + u\, h(x)$$
is simple, modular, and source-data-free at adaptation time.
2. **Theory:**
   - fallback guarantee:
$$\widetilde{O}\left(n^{-2\beta/(2\beta+d)} + \frac{p}{n}\right)$$
   - adaptive guarantee:
$$\widetilde{O}\left(\rho^{*\,2d/(2\beta+d)} n^{-2\beta/(2\beta+d)} + \frac{p}{n}\right)$$
   - formal no-negative-transfer corollary over $\mathcal{F}^\beta(f_{\mathrm{rep}}, \gamma)$
3. **Experiments:** robust under shift, noise, semantic perturbation, imbalance, and uniquely supports adapt-time multimodality extension.