

000 001 002 003 004 005 EXPLORING REPRESENTATION INVARIANCE IN FINE- 006 TUNING 007 008 009

010 **Anonymous authors**
011 Paper under double-blind review
012
013
014
015
016
017
018
019
020
021
022

023 ABSTRACT 024

025 Foundation models pretrained on large-scale natural images are widely adapted to
026 various cross-domain low-resource downstream tasks, benefiting from generaliz-
027 able and transferable patterns captured by their representations. However, these
028 representations are later found to gradually vanish during finetuning, accompanied
029 by a degradation of model’s original generalizability. In this paper, we argue
030 that such tasks can be effectively adapted without sacrificing the benefits of pre-
031 trained representations. We approach this by introducing *Representation Invariance*
032 *FineTuning (RIFT)*, a regularization that maximizes the representation similarity
033 between pretrained and finetuned models by leveraging orthogonal invariance of
034 manifolds in a computationally efficient way. Experiments demonstrate that our
035 method is compatible with mainstream finetuning methods, offering competitive or
036 even enhanced performance and better preservation of the generalizability.
037
038

039 1 INTRODUCTION 040

041 Foundation models pretrained on large-scale natural images have been widely recognized as strong
042 initializations for various downstream tasks Siméoni et al. (2025); Yu et al. (2025); Fang et al. (2023),
043 particularly in data-scarce scenarios Liu et al. (2024a); Zhang et al. (2024a;b), as low-resource
044 learning tasks usually fail to train a powerful model to cover the complex data distributions, resulting
045 in poor performance Tian et al. (2020). A representative example is medical image analysis Wang
046 et al. (2023); Dai et al. (2023), where collecting labeled samples is difficult due to privacy concerns,
047 rare diseases, heterogeneous sources, and the high annotation cost. In contrast, the generalizable
048 and transferable patterns captured by pretrained representations can strongly compensate for this
049 limitation, which are shown to be shared across tasks Mehra et al. (2024).

050 Meanwhile, such tasks typically exhibit significant domain gaps, which necessitate finetuning (*i.e.*,
051 training on specific datasets with smaller learning rates, fewer epochs and selective parameters) to
052 successfully adapt foundation models Jia et al. (2022b); Chen et al. (2022); Hu et al. (2022). The
053 primary intention of finetuning is to implicitly minimize Euclidean distance shift of pretrained model
054 in parameter space, thereby not only enabling better transfer of established semantic knowledge
055 to downstream tasks but also preserving the inherent generalizability to support future incremental
056 requirements, such as continual and multi-task learning Wang et al. (2024); Zhang & Yang (2021).

057 However, these benefits of pretrained representations are not retained as expected, and instead degrade
058 severely during finetuning Wang et al. (2025); Chen et al. (2025); Kumar et al. (2022). Particularly,
059 recent Platonic Representation Hypothesis reveals that foundation models possess enough capacity
060 and scalability to capture non-conflicting shared representations across modalities and tasks, yet this
061 potential is often wasted, as simplicity bias may converge along shortcut paths tailored to the current
062 task Huh et al. (2024). This leaves us wondering, *how can pretrained representations avoid being*
063 *shaved away by such an Occam’s razor in cross-domain low-resource finetuning?*

064 Many biological systems adjust to new environments while protecting their core functions, a stability
065 that arises from constraints imposed throughout the evolution, *e.g.*, protein structures influenced by
066 hydrophobic interactions remain robust in their overall fold even when large temperature fluctuations
067 perturb local residues Hatakeyama & Kaneko (2015); Tang et al. (2020). Inspired by this principle, we
068 hypothesize that pretrained and finetuned representations fail to coexist due to insufficient constraints.
069

054 Intuitively, one can restrict finetuning by aligning it with pretrained representations to prevent
 055 excessive drift. This alignment can be achieved through similarity measures, with Centered Kernel
 056 Alignment (CKA) as a popular choice Kornblith et al. (2019). A straightforward approach is to
 057 add a regularization term that keeps representation similarity above certain threshold to the loss
 058 function, but it incurs considerable computational cost due to the high complexity of pairwise CKA
 059 calculations. To alleviate this, we propose *Represantaion Invariance FineTuning (RIFT)* to make three
 060 efforts. First, we exploit the orthogonal invariance of CKA by maintaining orthogonality between
 061 two feature embeddings, as a cheaper alternative. Second, we perform distributional rather than
 062 sample-wise alignment for each mini-batch, leveraging the precomputed mean and covariance for an
 063 efficient statistical approximation. Third, we use only feature embeddings from the last layer before
 064 task head to avoid orthogonality violations in multi-layer architectures or nonlinear activations.
 065

066 Overall, our contributions can be summarized as follows:

- 067 • We demonstrate that the benefits of pretrained representations can be well preserved while still
 068 effectively adapting to downstream tasks, especially in cross-domain low-resource scenarios.
 069 • We propose **RIFT**, a simple regularization for finetuning to constrain representation similarity by
 070 the orthogonal invariance of CKA with improved computational efficiency.
 071 • Our method is compatible with mainstream finetuning methods, achieving competitive or even
 072 enhanced performance while better preserve the generalizability.

073 We hope our work can shed some light on the finetuning paradigms that emphasize both generalization
 074 and adaptation.

077 2 RELATED WORK

079 2.1 REPRESENTATION INVARIANCE AND SIMILARITY

081 Pretrained representations have been shown to capture rich and diverse features from real-world
 082 datasets, and are widely used to accelerate and stabilize the convergence of downstream tasks Yu
 083 et al. (2025); Wu et al. (2023); He et al. (2022); Liang et al. (2025). However, these representations
 084 have been found to inevitably degrade during finetuning, with downstream task performance being
 085 adversely affected and catastrophic forgetting also weakening pretrained semantic knowledge. Con-
 086 sequently, several studies have proposed methods to address these issues Aghajanyan et al. (2021);
 087 Razdaibiedina et al. (2023); Ma et al. (2021). Yet, despite these efforts, the generalizability of
 088 pretrained representation is neglected, and remains unverified as inevitable degradation, with no
 089 effective approaches yet established Wang et al. (2025); Chen et al. (2025); Kumar et al. (2022). In
 090 this paper, we confront the challenge of whether pretrained representations and their generalizability
 091 can be simultaneously preserved, *i.e.*, exploring the representation invariance in finetuning.

092 Representation similarity metrics provide a natural tool to quantify such invariance. Different
 093 similarity measures Klabunde et al. (2025); Huh et al. (2024) have been proposed to compare
 094 representations across layers or models to better understand neural network behaviors, *e.g.*, Canonical
 095 Correlation Analysis (CCA) Morcos et al. (2018) and Centered Kernel Alignment (CKA) Kornblith
 096 et al. (2019). These metrics exhibit various desirable invariances, such as invariance to invertible
 097 linear transformations, orthogonal transformations, and isotropic scaling. Among them, CKA offers
 098 superior consistency across architectures, stronger invariance, and more interpretable results.

099 2.2 FINETUNING AND ORTHOGONAL CONSTRAINTS

101 Typically, finetuning is divided into full finetuning Lv et al. (2024) and parameter-efficient finetuning
 102 (PEFT) Ding et al. (2023). Full finetuning retrains all pretrained parameters Kirillov et al. (2023);
 103 Touvron et al. (2023); Liu et al. (2023), while PEFT updates only a subset or newly introduced ones,
 104 *e.g.*, prompt tuning Zu et al. (2024); Jia et al. (2022b); Bahng et al. (2022), adapter tuning Chen et al.
 105 (2022); He et al. (2021); Sung et al. (2022), and LoRA Hu et al. (2022); Liu et al. (2024b); Hayou
 106 et al. (2024).

107 Introducing orthogonal constraints during training and finetuning Qiu et al. (2023); Liu et al. (2024c);
 108 Ma et al. (2024); Yang et al. (2025); Qiu et al. (2025); Duan et al. (2025) has been proposed to

108 preserve pretrained generative abilities. The motivation lies in the fact that orthogonal transformations
 109 preserve both the spectral norm of weight matrices and angular relationships Qiu et al. (2023; 2025).
 110 In particular, OFT Qiu et al. (2023) enforces such constraints on the weight matrix to maintain
 111 hyperspherical energy. *In contrast, RIFT is fundamentally different from these approaches: (1)*
 112 *RIFT operates directly on the final representations rather than imposing constraints on specific*
 113 *weights, thereby avoiding the limitations detailed in Sec. 4.3. (2) RIFT prioritizes preserving the*
 114 *generalizability of pretrained representations, rather than solely transferring established semantic*
 115 *knowledge to downstream tasks, though it achieves this as well.*

117 3 PRELIMINARY: CKA FOR REPRESENTATION SIMILARITY

119 **Notation.** Bold symbols denote matrices or vectors (e.g., \mathbf{X} and \mathbf{x}), while scalars are written in
 120 lowercase (e.g., x). The dataset $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^{n \times d}$ where n is the number of samples and d is
 121 feature dimension. The Frobenius norm is denoted by $\|\cdot\|_F$.

122 **Model.** A model is represented by f_θ , where θ denotes the parameters, and we simplify it as f . The
 123 notation $\mathbf{F}_\theta(\mathbf{X})$ denotes features extracted by backbone, which is distinct from the classification
 124 output $f_\theta(\mathbf{X})$. They are omitted as \mathbf{X} and \mathbf{F}_θ . The parameters of pretrained model are denoted as θ_0 .

125 CKA has been a widely used metric for representation similarity. Formally, the centered feature
 126 embedding $\mathbf{F}_{\theta,c} \in \mathbb{R}^{n \times d}$ computed by f_θ on \mathbf{X} is defined as

$$128 \quad \mathbf{F}_{\theta,c} := \mathbf{F}_\theta - \frac{1}{n} \mathbf{1} \mathbf{1}^\top \mathbf{F}_\theta \quad (1)$$

129 where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. The linear CKA of \mathbf{F} between f_{θ_0} and f_θ is then given by

$$131 \quad \text{CKA}(\mathbf{F}_{\theta_0}, \mathbf{F}_\theta) = \frac{\|\mathbf{F}_{\theta_0,c}^\top \mathbf{F}_{\theta,c}\|_F^2}{\|\mathbf{F}_{\theta,c}^\top \mathbf{F}_{\theta,c}\|_F \|\mathbf{F}_{\theta_0,c}^\top \mathbf{F}_{\theta_0,c}\|_F} \quad (2)$$

134 which is the cosine similarity of centered features' Gram matrices and quantifies the structural
 135 alignment degree. It is particularly sensitive to dimensional collapse, while remaining invariant to
 136 orthogonal transformations and isotropic scalings. The invariance is stated as

137 **Property 3.1** (CKA Similarity-Transformation Invariance). *For any scalar $\alpha > 0$ and orthogonal*
 138 *matrix $\mathbf{Q} \in O(d)$, $\mathbf{F}_\theta = \alpha \mathbf{F}_{\theta_0} \mathbf{Q}$ satisfies*

$$139 \quad \text{CKA}(\mathbf{F}_{\theta_0}, \mathbf{F}_\theta) = 1 \quad (3)$$

141 where such transformations ensure a high CKA value, implying that feature embeddings remain
 142 highly similar. This condition motivates the following definition.

143 **Definition 3.2** (Similarity-Invariant Parameter Subspace). *Given θ_0 , define the set*

$$144 \quad \mathcal{M}_{\theta_0} := \{\theta \in \Theta \mid \exists \mathbf{Q} \in O(d), \exists \alpha > 0 \text{ such that } \mathbf{F}_\theta = \alpha \mathbf{F}_{\theta_0} \mathbf{Q}\}. \quad (4)$$

146 which offers a promising way to identify finetuned models preserving the pretrained representation,
 147 and we primarily focus on feature embeddings from the last layer before task head.

149 4 METHOD

151 4.1 PROBLEM FORMULATION

153 The goal is to find a finetuned model that both maximizes downstream task performance and preserves
 154 representation similarity with pretrained model. This is formulated as an optimization problem:

$$155 \quad \theta^* = \arg \min_{\theta} \mathcal{L}_{\text{cls}}(\theta) \quad \text{s.t.} \quad \text{CKA}(\mathbf{F}_{\theta_0}, \mathbf{F}_\theta) \geq \epsilon \quad (5)$$

156 where $\mathcal{L}_{\text{cls}}(\theta)$ is the classification loss and $\epsilon \in [0, 1]$ is threshold. To solve Eq. 5, we can introduce
 157 the *Representation Similarity Constrained (RSC)* Loss and try to minimize it:

$$158 \quad \mathcal{L}_{\text{RSC}}(\theta) = \mathcal{L}_{\text{cls}}(\theta) + \lambda \mathcal{L}_{\text{CKA}}(\theta) \quad (6)$$

$$159 \quad \mathcal{L}_{\text{CKA}}(\theta) = 1 - \text{CKA}(\mathbf{F}_{\theta_0}, \mathbf{F}_\theta) \quad (7)$$

161 where $\mathcal{L}_{\text{CKA}}(\theta)$ is the similarity loss, and λ is the regularization strength. However, computing
 $\mathcal{L}_{\text{CKA}}(\theta)$ causes considerable computational overhead, as pairwise CKA incurs high complexity.

162 4.2 REPRESENTATION INVARIANCE FINETUNING
163

164 To efficiently preserve representation similarity, we
165 reformulate Eq. 5 with Definition 3.2 as

$$166 \theta^* = \arg \min_{\theta \in \mathcal{M}_{\theta_0}} \mathcal{L}(\theta) \quad (8)$$

168 which implicitly guarantees representation similarity,
169 and replaces the explicit CKA computation. Given a
170 learnable orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$ and scaling
171 factor α , we enforce the finetuned representation to
172 reside on orthogonal manifold of pretrained represen-
173 tation with Property 3.1 by

$$174 \|\mathbf{F}_\theta(\mathbf{X}) - \alpha \mathbf{F}_{\theta_0}(\mathbf{X}) \mathbf{Q}\|_F^2 \quad (9)$$

175 Notably, directly imposing this constraint either de-
176 demands a sample-wise forward pass through the pre-
177 trained model or substantial storage. We instead ap-
178 proximate it by matching the mini-batch $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$ of for every mini-batch:

$$179 \|\mu_\theta - \alpha \mu_{\theta_0} \mathbf{Q}\|_F^2 \quad \text{and} \quad \|\Sigma_\theta - \alpha^2 \mathbf{Q}^\top \Sigma_{\theta_0} \mathbf{Q}\|_F^2 \quad (10)$$

180 The analysis of this relaxation rationality is provided in Theorem A3.5. In practice, we find that
181 the transformation $\alpha \mathbf{Q}$ has already adjusted both scale and orientation of pretrained representations,
182 implicitly aligning their mean with current representations. We also observe that dynamic α brings
183 little convergence benefits and increases training burden. Therefore, we simplify Eq. 10 to covariance
184 matching with $\alpha = 1$, and further propose the *Representation Invariance FineTuning (RIFT)* Loss:

$$185 \mathcal{L}_{\text{RIFT}}(\theta, \mathbf{Q}) = \mathcal{L}_{\text{cls}}(\theta) + \mathcal{L}_{\text{Cov}}(\theta, \mathbf{Q}) \quad (11)$$

$$186 \mathcal{L}_{\text{Cov}}(\theta, \mathbf{Q}) = \|\Sigma_\theta - \mathbf{Q}^\top \Sigma_{\theta_0} \mathbf{Q}\|_F^2 \quad (12)$$

188 **Generalization and Adaptation.** Here, \mathbf{Q} applies only
189 global isometric rotations to the whole feature space with-
190 out altering relative angles among different class clusters
191 or sample vectors, thereby preserving semantic consist-
192 ency and generalizability of the pretrained representa-
193 tion, *i.e.*, any sample, whether in- or out-of-pretrained-
194 distribution, are mapped to its original predicted label
195 after the current orthogonal transformation. With this
196 foundation, $\mathcal{L}_{\text{cls}}(\theta)$ promotes the representation to gen-
197 erate additional semantic structures for finetuning datasets.

198 **Compatibility and Efficiency.** Importantly, RIFT reveals
199 that generalization and adaptation are not mutually exclu-
200 sive, and can serve as a plug-and-play training strategy
201 compatible with mainstream finetuning methods, since it
202 constrains only the output representations without modi-
203 fying networks. RIFT can also be easily integrated with
204 more advanced orthogonal finetuning techniques by up-
205 dating \mathbf{Q} . Through trace-based decoupling of covariance
206 from individual samples, the computational complexity of
207 RIFT is reduced from $O(nd^2)$ to $O(d^2)$. Detailed training time is presented in Tab. A5.

208 4.3 WHY ORTHOGONAL CONSTRAINTS IMPOSED ONLY AT FINAL FEATURE EMBEDDINGS
209

210 An intriguing question is why we apply the orthogonal transformation to last-layer feature embeddings
211 before task head rather than output of another or each layer. The reason initially lies in fact that final
212 feature embeddings are used for task decisions, which must be kept. Although imposing individual
213 constraints on a single linear layer guarantees similar outputs, extending it to multi-layer models or
214 introducing nonlinearities (*e.g.*, activations) breaks the full model representation invariance.

215 We first expand the network depth and apply orthogonal transformations on just one layer, using
an autoencoder trained for image reconstruction in Fig 2. As shown in Fig. 2 (b), applying random

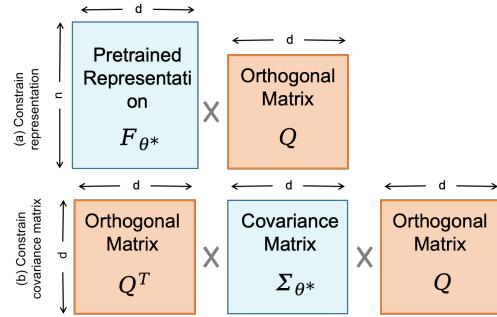


Figure 1: (a) Orthogonal transformation of the pretrained representation. (b) Orthogonal transformation of the covariance.

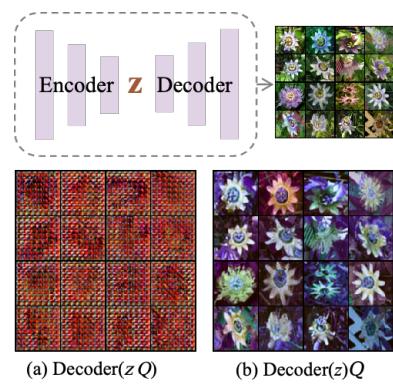


Figure 2: (a) Applying Orthogonal transformation at the intermediate layer. (b) Applying orthogonal transformation at the last layer.

orthogonal transformations to the final feature embeddings of pretrained model allows for clear image reconstruction. In contrast, results in Fig 2 (a) fail because constraint is only added to the encoder, while the decoder remains untransformed, causing a distribution mismatch and thus hindering the complete recovery of feature embeddings in deeper layers. This demonstrates that applying constraints at other layers cannot guarantee the invariance of final feature embeddings.

Then we further apply layer-wise random independent orthogonal transformations, and conduct a toy experiment in Fig. 3 with two Gaussian-distributed classes on 6-layer linear or nonlinear networks. The results indicate that they still results in severe representation degradation in deep networks with nonlinearities further making the invariance preservation uncertain. These findings highlight constraining final feature embeddings is the key to global representation invariance.

5 EVALUATION

5.1 SETTING

We choose medical image classification as the representative cross-domain low-resource downstream task. The backbone is a Vision Transformer Dosovitskiy et al. (2020) (ViT-B/16), pretrained on ImageNet-1K (12 encoder layers, 768-dim embeddings, 12 attention heads). We compare our method RIFT against RSC by integrating them into mainstream finetuning methods, including full finetuning (FULL), classification-head only (LINEAR), training from scratch (SCRATCH), and PEFT(VPT, AdaptFormer, LoRA). Evaluation metrics contains CKA for representation similarity measure (Sim) as well as accuracy (Acc) and mean Average Precision (mAP) for classification performance. For reference, the representation similarity of LINEAR is always 1 since backbone is frozen. Experiments are conducted on five medical datasets, including three MedFM subsets Wang et al. (2023) (Chest, Colon, Endo), ISIC2018 Codella et al. (2019), and APTOS2019 Karthik & Dane. They covers diverse modalities (skin, fundus, chest X-ray, pathology, colonoscopy) and tasks (multi-class, binary, multi-label). Additional training and evaluation details are provided in Appendix A1.

5.2 MAIN RESULT

5.2.1 FINETUNING RESULT

Quantitative result on medical image classification. The results in Table 1 demonstrate that RIFT consistently improves representation similarity across both single-label and multi-label datasets, while maintaining competitive and even enhanced accuracy. For non-PEFT methods, although FULL finetuning achieves higher Acc than SCRATCH, it often results in lower Sim, highlighting the degradation of pretrained representation with FULL. In contrast, both RIFT and RIFT* are able to preserve or even enhance Sim compared to FULL, with RIFT* achieving the best balance between downstream task performance and pretrained representation preservation across datasets. Notably, on single-label datasets such as ISIC2018 and APTOS2019, RIFT variants maintain Acc close to FULL, while substantially improving Sim by 50–87.5%, indicating that the learned patterns remain closer to pretrained feature space, which is critical for transferability and generalization. Compared to RSC, RIFT achieves higher representation similarity with better downstream adaptation, suggesting that orthogonal transformations provide a more effective constraint than direct CKA-based regularization.

For PEFT methods, including LoRA, Adaptformer, and VPT, the trends are similar but more pronounced. While these methods already restrict parameter updates to preserve pretrained knowledge, applying RIFT still yields substantial gains in Sim, often exceeding 15–28%, while the Acc remains comparable or slightly higher in RIFT* variants. This demonstrates that RIFT effectively regularizes the adaptation process without sacrificing performance, leading to representations both task-effective and semantically faithful to the original model. Moreover, across multi-label datasets, where maintaining correlations between multiple outputs is crucial, RIFT significantly enhances Sim, suggesting

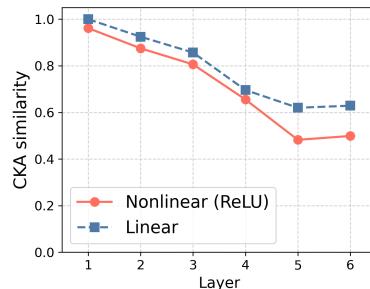
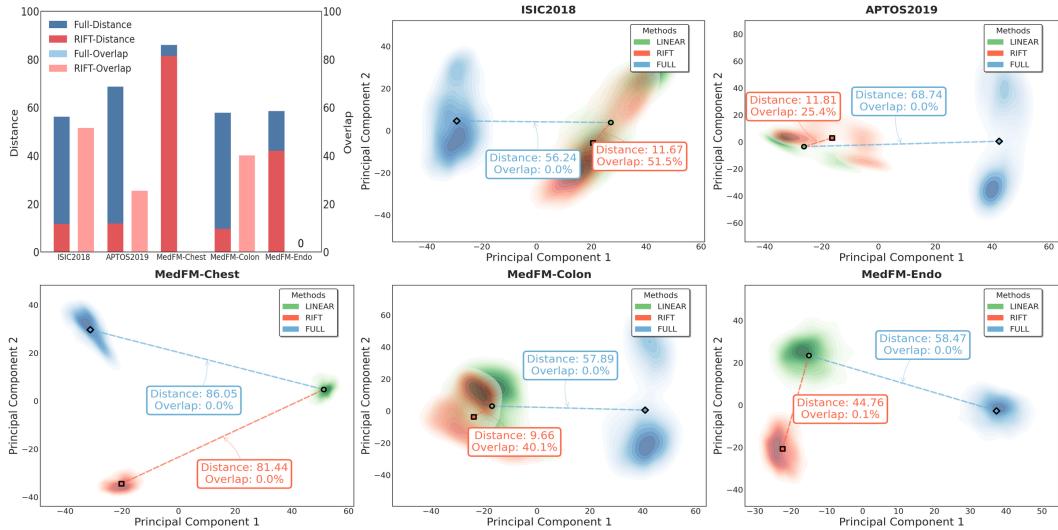


Figure 3: Representation similarity diminishes with layer-wise orthogonal constraints.

270
 271 **Table 1: Quantitative result on medical image classification.** For each task, the average result
 272 from 3 runs is reported. Dataset type is explicitly indicated: single-label (left) vs. multi-label (right).
 273 Methods are grouped into non-PEFT and PEFT for clarity. **Bold** indicates the best performance.
 274 Values in the last column show relative changes (%): **+-** denote increase/decrease relative to the first
 275 row of each block. RIFT refers to the default setting with regularization coefficient $\lambda = 1$, while
 276 RIFT* and RSC* correspond to the best λ selected for each dataset.

Model	Single-label datasets						Multi-label datasets			Average		
	ISIC2018 (7)		APTO2019 (5)		MedFM-Colon (2)		MedFM-Chest (19)		MedFM-Endo (4)		Acc/mAP	Sim
	Acc	Sim	Acc	Sim	Acc	Sim	mAP	Sim	mAP	Sim		
Non-PEFT Methods												
SCRATCH	66.91	0.36	69.49	0.37	89.42	0.67	11.89	0.47	17.08	0.39	50.96	0.45
FULL	84.63	0.32	84.06	0.52	99.74	0.58	33.56	0.58	58.11	0.61	72.02	0.52
+RSC*	85.19	0.35	84.97	0.59	99.51	0.82	34.62	0.44	57.52	0.67	+0.47%	+9.62%
+RIFT (ours)	83.95	0.60	83.61	0.73	99.05	0.82	33.11	0.55	55.42	0.70	-1.37%	+30.77%
+RIFT*(ours)	84.85	0.48	85.25	0.54	99.51	0.84	35.24	0.63	58.62	0.66	+0.93%	+21.15%
PEFT Methods												
LINEAR	73.57	1.00	77.87	1.00	94.24	1.00	23.55	1.00	37.80	1.00	61.41	1.00
LoRA Hu et al. (2022)	80.85	0.38	81.01	0.61	96.17	0.78	25.91	0.59	39.56	0.77	64.70	0.63
+RIFT(ours)	77.51	0.74	79.24	0.79	96.07	0.82	24.28	0.63	39.24	0.76	-2.22%	+20.18%
+RIFT*(ours)	80.92	0.40	81.01	0.61	96.37	0.79	25.77	0.61	39.88	0.77	+0.14%	+1.59%
Adaptformer Chen et al. (2022)	80.42	0.51	80.60	0.75	97.25	0.78	25.13	0.61	40.08	0.67	64.69	0.66
+RIFT(ours)	77.65	0.74	80.15	0.85	95.87	0.86	23.82	0.59	38.63	0.75	-2.27%	+15.15%
+RIFT*(ours)	79.96	0.56	80.60	0.81	97.54	0.81	25.24	0.65	40.55	0.71	+0.14%	+7.58%
VPT Jia et al. (2022a)	77.71	0.46	78.96	0.60	94.79	0.72	24.00	0.51	41.64	0.65	63.42	0.59
+RIFT(ours)	76.32	0.74	78.69	0.85	95.48	0.86	22.19	0.59	37.70	0.75	-2.11%	+28.81%
+RIFT*(ours)	78.31	0.50	78.96	0.66	95.48	0.86	24.58	0.47	40.54	0.67	+0.24%	+4.43%



311 **Figure 4: PCA distribution visualization.** We compare first two principal components of LINEAR,
 312 FULL, and FULL+RIFT($\lambda = 1$) features across five medical image datasets. The first figure
 313 summarizes the distance and overlap of the pretrained model features. Darker colors indicate higher
 314 feature density, while lighter colors indicate lower density.

315
 316 better preservation of semantic structure. By coupling adaptation with representation similarity, RIFT
 317 preserves established semantic knowledge, allowing it to serve downstream tasks more effectively.

318
 319 **PCA distribution visualization.** Fig. 4 shows the PCA feature distribution results of samples
 320 extracted from pretrained and finetuned models. The bar charts summarize the distance from the
 321 pretrained model’s feature center to the finetuned model’s feature center, as well as the overlap
 322 between features. FULL shows a noticeable distance and lack of overlap between features, indicating
 323 significant representation differences compared to LINEAR. In contrast, RIFT exhibits a high
 324 overlap of pretrained features on the ISIC2018, APTOS2019, and MedFM-Colon datasets. Even

for the multi-label classification datasets, MedFM-Chest and MedFM-Endo, although the expected overlap is not observed, the distance is closer compared to FULL. These results indicate that pretrained representations can be preserved without damaging their semantic structure by imposing orthogonal constraints, upon which new branches for downstream tasks can grow. This may unlock a temporal view of the Platonic Representation Hypothesis. This may unlock a temporal view of the Platonic Representation Hypothesis, suggesting that models have sufficient capacity to accommodate pretrained and finetuned knowledge simultaneously, yielding a shared representation.

Table 2: **Quantitative result on different backbones.** We adopt large-scale ViT pretrained on ImageNet-21K with supervision and DINOV2 ViT-base.

Model	Single-label datasets						Multi-label datasets				Average	
	ISIC2018 (7)		APTOPS2019 (5)		MedFM-Colon (2)		MedFM-Chest (19)		MedFM-Endo (4)		Acc/mAP	Sim
	Acc	Sim	Acc	Sim	Acc	Sim	mAP	Sim	mAP	Sim		
ViT-large(ImageNet-21K)												
Scratch	66.07	0.46	69.67	0.40	89.88	0.59	12.12	0.47	17.30	0.34	51.01	0.45
Linear	78.44	1.00	78.69	1.00	95.19	1.00	25.13	1.00	38.11	1.00	63.11	1.00
FULL	83.33	0.56	83.61	0.88	99.71	0.69	35.39	0.53	55.58	0.67	71.52	0.67
+RIFT (ours)	83.66	0.60	82.24	0.89	99.12	0.73	34.39	0.81	58.23	0.66	+0.01%	+11.05%
+RIFT* (ours)	84.13	0.52	83.06	0.87	99.51	0.75	38.06	0.67	58.23	0.66	+1.50%	+4.56%
ViT-base(DINOV2 Oquab et al. (2023))												
Scratch	60.58	0.22	54.37	0.65	85.36	0.69	10.46	0.78	16.08	0.37	45.37	0.54
Linear	75.53	1.00	80.33	1.00	94.06	1.00	23.24	1.00	36.12	1.00	61.85	1.00
FULL	76.82	0.30	76.78	0.49	98.13	0.53	12.29	0.64	22.20	0.37	57.24	0.47
+RIFT (ours)	69.78	0.40	76.50	0.52	92.34	0.68	12.83	0.62	38.18	0.35	+1.19%	+10.30%
+RIFT* (ours)	79.63	0.45	76.50	0.52	99.71	0.48	12.83	0.62	40.15	0.36	+7.89%	+4.39%

Quantitative result on different backbones. Tab. 2 presents the results of RIFT applied to ViT-large and DINOV2 backbones. When applied to larger models and the self-supervised DINOV2 backbone, RIFT consistently improves representation similarity while maintaining, or in some cases slightly improving, classification performance. Notably, larger models overall exhibit higher post-finetuning representation similarity, indicating stronger representational stability. In addition, the self-supervised pretrained DINOV2 model shows that finetuning only the linear head (LINEAR) outperforms full finetuning (FULL), especially on the MedFM-Chest dataset where the gap is substantial, highlighting the superior generalization ability of self-supervised pretraining.

5.3 GENERALIZATION RESULT

Table 3: **Qualitative result on zero-shot natural image classification.** We adopt model finetuned on ISIC2018 with several classical natural image datasets Parkhi et al. (2012); Nilsback & Zisserman (2008); Netzer et al. (2011); Krause et al. (2013); Wah et al. (2011) estimated using 20-NN. RIFT* indicates a regularization coefficient of $\lambda = 0.6$.

Model	Oxford-IIIT Pet	Oxford Flowers	SVHN	Stanford Cars	CUB-200	Avg
Pretrained Backbone	84.19	96.93	38.39	25.50	54.96	59.99
FULL (Sim=0.32)	82.03	90.41	38.36	26.03	55.10	58.39
+RIFT (Sim=0.60)	83.65	95.44	37.83	25.61	55.29	59.56
+RIFT* (Sim=0.48)	84.38	95.38	39.59	27.11	55.29	60.35

Qualitative result on zero-shot natural image classification. As shown in Table 3, we performed KNN evaluation on five unseen datasets to compare the generalization ability of different finetuned models. Supporting our hypothesis, the pretrained models generally exhibit better overall generalization performance than the model fully finetuned on the ISIC2018 dataset (FULL). Furthermore, compared to FULL, models with higher similarity (RIFT and RIFT*) tend to better preserve the generalization capability of the pretrained models. Interestingly, RIFT and RIFT* even improves generalization, likely due to incorporating additional transferable patterns in finetuning data.

Quantitative result on theoretical generalizability. Tab. 4 evaluates the generalizability of finetuned models by comparing the sharpness of the loss function (i.e., stability under perturbations),

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

Table 4: **Quantitative result on theoretical generalizability.** We validate sharpness $\max_{\|\epsilon\| \leq \rho} \mathcal{L}_{\text{cls}}(\theta + \epsilon) - \mathcal{L}_{\text{cls}}(\theta)$ with lower values indicating better generalization across methods: FULL, RIFT($\lambda = 1$), and LINEAR, with $\rho = 0.01$.

Metric	Model	ISIC2018	APTO2019	MedFM-Chest	MedFM-Colon	MedFM-Endo	Avg.
$\max_{\ \epsilon\ \leq \rho} \mathcal{L}_{\text{cls}}(\theta + \epsilon) \downarrow$	LINEAR	0.7292	0.6442	4.7553	0.1995	1.1375	1.4931
	FULL	0.6398	0.5701	4.3836	0.0509	1.0859	1.3461
	RIFT (ours)	0.5558	0.5262	4.3338	0.0677	1.0036	1.2974
$\mathcal{L}_{\text{cls}}(\theta) \downarrow$	LINEAR	0.7159	0.5989	4.7415	0.1841	1.1343	1.4749
	FULL	0.6125	0.5049	4.3727	0.0305	1.0743	1.3190
	RIFT (ours)	0.5312	0.5082	4.3236	0.0496	0.9957	1.2817
Sharpness \downarrow	LINEAR	0.0133	0.0453	0.0138	0.0154	0.0032	0.0182
	FULL	0.0273	0.0652	0.0109	0.0204	0.0116	0.0271
	RIFT (ours)	0.0246	0.0180	0.0102	0.0181	0.0080	0.0158

measured as $\max_{\|\epsilon\| \leq \rho} \mathcal{L}_{\text{cls}}(\theta + \epsilon) - \mathcal{L}_{\text{cls}}(\theta)$, across different finetuning methods. Lower sharpness reflects better generalizability. LINEAR outperforms FULL with an average sharpness of 0.0182, demonstrating its robustness due to the pretrained initialization. FULL exhibits a higher average sharpness of 0.0271, indicating weaker generalization. RIFT achieves the lowest average sharpness of 0.0158, demonstrating its ability to achieve flatter minima and better generalization. Notably, RIFT consistently performs better than FULL across all datasets, with particularly strong results on APTOS2019, where its sharpness (0.0180) is significantly lower than that of FULL (0.0652). These results highlight that, under representation similarity constraints, RIFT actively guides the optimization landscape toward flatter and more generalizable minima.

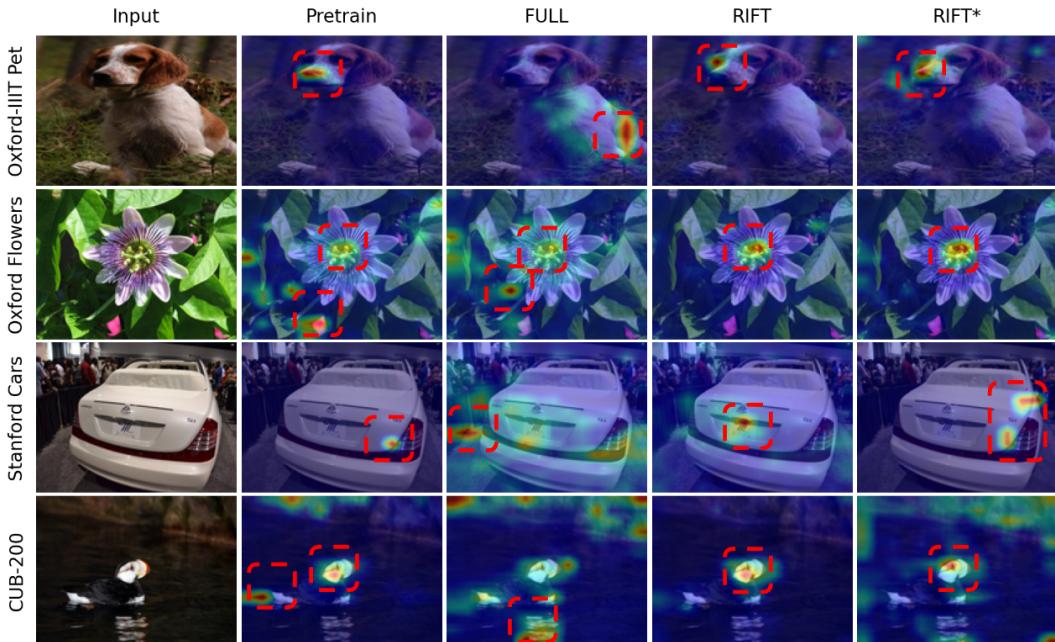


Figure 5: **Attention heatmap visualization.** We give more qualitative results on zero-shot natural image classification to further demonstrate the generalizability of RIFT and RIFT*. The red boxes highlight the regions most attended to by each method. Images are taken from previously unseen datasets: Oxford-IIIT Pet, Oxford Flowers, Stanford Cars, and CUB-200.

Attention heatmap visualization. As shown in Fig. 5, the pretrained model focuses on the nose and mouth regions of dogs in Oxford-IIIT Pet and on the head regions of birds in CUB-200. In contrast, after direct finetuning (FULL), the attention shifts to less relevant regions, such as the back of the dog or the bird's reflection in the water. This supports the observation that pretrained and finetuned (FULL) models generalize differently on unseen datasets, although both tend to overlook the flower centers in the Oxford Flowers dataset. RIFT and RIFT, however, retain the generalization patterns of

432 pretrained representations, correctly attending to class-relevant objects across all cases, and in some
 433 instances even improving upon them (e.g., on the flower dataset), thereby further strengthening the
 434 generalization ability of current learned representations.

436 5.4 ABLATION STUDY

438 Ablation studies are conducted to further explore the impact of different components. The key
 439 differences among the examined methods lie in whether the Q matrix is learnable, the application of
 440 mean alignment (*i.e.*, using μ), the chosen batch size and the regularization coefficient λ .

441 **Ablation of matrix Q .** As shown in Tab. 5 (a), employing a learnable Q matrix improves accuracy
 442 and similarity metrics by 0.4% and 0.04, respectively, compared to using a fixed Q parameter.
 443 This demonstrates that a learnable orthogonal matrix can more effectively preserve pretrained
 444 representations and leverage their benefits to strengthen downstream task adaptation.

445 Table 5: **Ablation study on fixed Q , aligning μ , and batch size.**

(a) Ablation of matrix Q .			(b) Ablation of aligning μ .			(c) Ablation of batch size.		
Matrix Q	Acc	Sim	mean μ	Acc	Sim	Batch Size	Acc	Sim
Fixed Q	83.55	0.56	w/ μ	83.49	0.51	4	81.42	0.21
Learnable Q	83.95↑	0.60↑	w/o μ	83.95↑	0.60↑	32	83.33	0.30

453 **Ablation of alignment of μ .** Tab. 5 (b) shows that removing mean alignment (w/o μ) achieves better
 454 accuracy and similarity, with gains of 0.46% and 0.09 over applying it (w/ μ). This indicates that
 455 mean alignment may impose unnecessary constraints on feature distribution, limiting the model’s
 456 ability to exploit data structure. Without it, the model learns more flexible representations, yielding
 457 improved performance.

458 **Ablation of batch size.** Increasing
 459 the batch size from 4 to 128, as shown
 460 in Tab. 5 (c), significantly boosts ac-
 461 curacy and similarity metrics, with
 462 improvements of 2.53% and 0.39, re-
 463 spectively. A larger batch size pro-
 464 vides a more accurate estimation of the covariance matrix. The observed trend underscores the
 465 importance of selecting an appropriate batch size.

466 **Ablation of λ .** As shown in Table 6, we examine the effect of the regularization coefficient λ
 467 on different datasets and finetuning methods, varying it from 0.1 to 1.0. Overall, choosing an
 468 appropriate λ yields dual benefits of adaptation and generalization. Even under extreme settings
 469 (*e.g.*, $\lambda = 1$), downstream performance shows only a slight drop without sacrificing the effective
 470 adaptation. Additional results are provided in Tab. A2 and Tab. A3 in the Appendix.

472 6 CONCLUSION

474 We propose *Representation Invariance FineTuning (RIFT)*, a simple and efficient constraint that
 475 preserves pretrained representations by enforcing orthogonal-invariant CKA similarity between
 476 pretrained and finetuned models. Our experiments across cross-domain and low-resource scenarios
 477 show that RIFT integrates seamlessly with mainstream finetuning methods, achieving competitive
 478 or even improved downstream task performance while mitigating the loss of generalization. These
 479 findings suggest that effective adaptation does not need to come at the cost of established semantic
 480 knowledge and generalizability, and highlight the value of explicitly preserving representation
 481 invariance during finetuning. We believe this work opens up promising directions for finetuning
 482 paradigms that emphasize the compatibility of adaptation and generalization in foundation models.

483 **Limitations.** Although our work demonstrates the effectiveness of RIFT in vision foundation models,
 484 the challenge of preserving pretrained representations extends beyond vision. Future research should
 485 investigate its applicability to other modalities (*e.g.*, large language and multimodal models) and
 486 broader cross-domain or low-resource tasks, such as embodied AI, mathematics, and programming.

486 ETHICS STATEMENT
487488 This work complies with the Code of Ethics. Our research does not involve human subjects, sensitive
489 personal data, or experiments that could raise ethical concerns. No hical issues are associated with
490 the methods, experiments, or results presented in this paper.
491492 REPRODUCIBILITY STATEMENT
493494 Hyperparameter configurations are detailed in Section 5 and Appendix A1. We also intend to release
495 the code and model checkpoints to enable full reproduction of the results reported in this paper.
496497 REFERENCES
498499 Mmclassification. <https://github.com/open-mmlab/mmclassification>.
500501 Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and
502 Sonal Gupta. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=QQ08SN70M1V>.
503504 Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts
505 for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.
506507 Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu.
508 Understanding pre-training and fine-tuning from loss landscape perspectives. *arXiv preprint*
509 *arXiv:2505.17646*, 2025.
510511 Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo.
512 Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural*
513 *Information Processing Systems*, 35:16664–16678, 2022.
514515 Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David
516 Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion
517 analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging
518 collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.
519520 Zhiyong Dai, Jianjun Yi, Lei Yan, Qingwen Xu, Liang Hu, Qi Zhang, Jiahui Li, and Guoqiang Wang.
521 Pfemed: Few-shot medical image classification using prior guided feature enhancement. *Pattern*
522 *Recognition*, 134:109108, 2023.
523524 Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin
525 Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained
526 language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.
527528 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
529 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An
530 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*
531 *arXiv:2010.11929*, 2020.
532533 Haoran Duan, Shuai Shao, Bing Zhai, Tejal Shah, Jungong Han, and Rajiv Ranjan. Parameter
534 efficient fine-tuning for multi-modal generative vision models with möbius-inspired transformation.
535 *International Journal of Computer Vision*, 133(7):4590–4603, 2025.
536537 Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong
538 Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale.
539 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp.
540 19358–19369, 2023.
541542 Tetsuhiro S Hatakeyama and Kunihiko Kaneko. Reciprocity between robustness of period and
543 plasticity of phase in biological clocks. *Physical Review Letters*, 115(21):218101, 2015.
544

- 540 Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. In
 541 *International Conference on Machine Learning*, pp. 17783–17806. PMLR, 2024.
 542
- 543 Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing,
 544 and Luo Si. On the effectiveness of adapter-based tuning for pretrained language model adaptation.
 545 In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and*
 546 *the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,
 547 pp. 2208–2222, 2021.
- 548 Ruiwei He, Shuyang Sun, Jihan Yang, Song Bai, and Xiaojuan Qi. Knowledge distillation as efficient
 549 pre-training: Faster convergence, higher data-efficiency, and better transferability. In *Proceedings*
 550 *of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9161–9171, 2022.
 551
- 552 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
 553 Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- 554 Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representa-
 555 tion hypothesis. In *Forty-first International Conference on Machine Learning*, 2024.
- 556
- 557 Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and
 558 Ser-Nam Lim. Visual prompt tuning. In *European conference on computer vision*, pp. 709–727.
 559 Springer, 2022a.
- 560
- 561 Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and
 562 Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727.
 563 Springer, 2022b.
- 564 Maggie Karthik and Sohier Dane. Aptos 2019 blindness detection (2019). *URL* <https://kaggle.com/competitions/aptos2019-blindness-detection>.
- 565
- 566
- 567 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete
 568 Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings*
 569 *of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- 570
- 571 Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of neural
 572 network models: A survey of functional and representational measures. *ACM Computing Surveys*,
 573 57(9):1–52, 2025.
- 574
- 575 Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural
 576 network representations revisited. In *International conference on machine learning*, pp. 3519–3529.
 577 PMLR, 2019.
- 578
- 579 Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained
 580 categorization. In *Proceedings of the IEEE international conference on computer vision workshops*,
 581 pp. 554–561, 2013.
- 582
- 583 Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-
 584 tuning can distort pretrained features and underperform out-of-distribution. In *International*
 585 *Conference on Learning Representations*, 2022. *URL* <https://openreview.net/forum?id=UYneFzXSJWh>.
- 586
- 587 Yiming Liang, Tianyu Zheng, Xinrun Du, Ge Zhang, Jiaheng Liu, Xingwei Qu, Wenqiang Zu,
 588 Xingrun Xing, Chujie Zheng, Lei Ma, et al. Aligning instruction tuning with pre-training. *arXiv*
 589 *preprint arXiv:2501.09368*, 2025.
- 590
- 591 Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and
 592 Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions*
 593 *on Geoscience and Remote Sensing*, 62:1–16, 2024a.
- 594
- 595 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in*
 596 *neural information processing systems*, 36:34892–34916, 2023.

- 594 Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-
 595 Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first*
 596 *International Conference on Machine Learning*, 2024b.
- 597 Weiyang Liu, Zeju Qiu, Yao Feng, Yuliang Xiu, Yuxuan Xue, Longhui Yu, Haiwen Feng, Zhen
 598 Liu, Juyeon Heo, Songyou Peng, et al. Parameter-efficient orthogonal finetuning via butterfly
 599 factorization. In *ICLR*, 2024c.
- 600 Kai Lv, Yuqing Yang, Tengxiao Liu, Qipeng Guo, and Xipeng Qiu. Full parameter fine-tuning for
 601 large language models with limited resources. In *Proceedings of the 62nd Annual Meeting of the*
 602 *Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8187–8198, 2024.
- 603 Xiaofei Ma, Cicero dos Santos, and Andrew O Arnold. Contrastive fine-tuning improves robustness
 604 for neural rankers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*
 605 2021, pp. 570–582, 2021.
- 606 Xinyu Ma, Xu Chu, Zhibang Yang, Yang Lin, Xin Gao, and Junfeng Zhao. Parameter efficient
 607 quasi-orthogonal fine-tuning via givens rotation. *arXiv preprint arXiv:2404.04316*, 2024.
- 608 Akshay Mehra, Yunbei Zhang, and Jihun Hamm. Understanding the transferability of representations
 609 via task-relatedness. *Advances in Neural Information Processing Systems*, 37:116513–116546,
 610 2024.
- 611 Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural
 612 networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.
- 613 Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al.
 614 Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep*
 615 *learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011.
- 616 Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number
 617 of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp.
 618 722–729. IEEE, 2008.
- 619 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov,
 620 Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning
 621 robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- 622 Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012*
 623 *IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.
- 624 Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller,
 625 and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances*
 626 *in Neural Information Processing Systems*, 36:79320–79362, 2023.
- 627 Zeju Qiu, Simon Buchholz, Tim Z Xiao, Maximilian Dax, Bernhard Schölkopf, and Weiyang
 628 Liu. Reparameterized Ilm training via orthogonal equivalence transformation. *arXiv preprint*
 629 *arXiv:2506.08001*, 2025.
- 630 Anastasia Razdaibiedina, Ashish Khetan, Zohar Karnin, Daniel Khashabi, and Vivek Madan. Rep-
 631 resentation projection invariance mitigates representation collapse. In *The 2023 Conference on*
 632 *Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=58jpJdPgKi>.
- 633 Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose,
 634 Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv*
 635 *preprint arXiv:2508.10104*, 2025.
- 636 Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vi-adapter: Parameter-efficient transfer learning for
 637 vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and*
 638 *pattern recognition*, pp. 5227–5237, 2022.
- 639 Qian-Yuan Tang, Tetsuhiro S Hatakeyama, and Kunihiko Kaneko. Functional sensitivity and muta-
 640 tional robustness of proteins. *Physical Review Research*, 2(3):033452, 2020.

- 648 Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking
 649 few-shot image classification: a good embedding is all you need? In *European conference on*
 650 *computer vision*, pp. 266–282. Springer, 2020.
- 651
- 652 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
 653 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
 654 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- 655 Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd
 656 birds-200-2011 dataset. 2011.
- 657
- 658 Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu
 659 Gao, Jun Shen, Junjun He, Tian Shen, et al. A real-world dataset and benchmark for foundation
 660 model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023.
- 661 Han Wang, Adam White, and Martha White. Fine-tuning without performance degradation. *arXiv*
 662 *preprint arXiv:2505.00913*, 2025.
- 663
- 664 Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning:
 665 Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*,
 666 46(8):5362–5383, 2024.
- 667
- 668 Chenwei Wu, Holden Lee, and Rong Ge. Connecting pre-trained language model and downstream
 669 task via properties of representation. *Advances in Neural Information Processing Systems*, 36:
 47216–47238, 2023.
- 670
- 671 Yiting Yang, Hao Luo, Yuan Sun, Qingsen Yan, Haokui Zhang, Wei Dong, Guoqing Wang, Peng
 672 Wang, Yang Yang, and Hengtao Shen. Efficient adaptation of pre-trained vision transformer
 673 underpinned by approximately orthogonal fine-tuning strategy. *arXiv preprint arXiv:2507.13260*,
 2025.
- 674
- 675 Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and
 676 Saining Xie. Representation alignment for generation: Training diffusion transformers is easier
 677 than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.
 678 URL <https://openreview.net/forum?id=DJSZGGZYVi>.
- 679
- 680 Dongmei Zhang, Chang Li, Renrui Zhang, Shenghao Xie, Wei Xue, Xiaodong Xie, and Shanghang
 681 Zhang. Fm-ov3d: Foundation model-based cross-modal knowledge blending for open-vocabulary
 682 3d detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp.
 16723–16731, 2024a.
- 683
- 684 Junjie Zhang, Chenjia Bai, Haoran He, Zhigang Wang, Bin Zhao, Xiu Li, and Xuelong Li. Sam-
 685 e: leveraging visual foundation model with sequence imitation for embodied manipulation. In
 686 *Proceedings of the 41st International Conference on Machine Learning*, pp. 58579–58598, 2024b.
- 687
- 688 Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and*
 689 *data engineering*, 34(12):5586–5609, 2021.
- 690
- 691 Wenqiang Zu, Shenghao Xie, Qing Zhao, Guoqi Li, and Lei Ma. Embedded prompt tuning: Towards
 692 enhanced calibration of pretrained models for medical images. *Medical Image Analysis*, 97:103258,
 2024.
- 692
- 693
- 694
- 695
- 696
- 697
- 698
- 699
- 700
- 701

APPENDIX

LLM USAGE STATEMENT

ChatGPT-5 are used only as general-purpose assistive tools, such as for language polishing and improving readability. No part of the research ideation, experimental design, or substantive writing is generated by LLMs. The authors take full responsibility for the content of this paper.

A1 EXPERIMENT SETTINGS

We use the ViT-B/16 model Dosovitskiy et al. (2020), with an input image size of 224×224 and a patch size of 16. The pretrained model is trained in a supervised manner on ImageNet-1K. In our experiments, the batch size is set to 128 to obtain a better estimation of the covariance matrix of the finetuned features, and the learning rate is set to 6×10^{-4} . The experiments were conducted on 4 NVIDIA A100 40G GPUs, with each training session running for 50 epochs. Our code is built upon the MMClassification framework Mmc. Tab. A1 presents detailed information about the experimental dataset.

For Tab. 2, we use the ViT-Large model pretrained on ImageNet-21K in a supervised manner and the DINOv2 (Base) model pretrained in a self-supervised manner. Their patch sizes are 16 and 14, input image sizes are 384×384 and 518×518 , and feature dimensions are 1024 and 768, respectively.

Table A1: Details of the dataset specifications.

Dataset	Modality	Task Type	Classes	Train	Test	Metric
ISIC2018 Codella et al. (2019)	Dermoscopy	Multiclass	7	10,015	1,512	Accuracy
APTOPS2019 Karthik & Dane	Fundus	Multiclass	5	2,930	366	Accuracy
MedFM-Chest Wang et al. (2023)	X-ray	Multi-label	19	2,140	3,869	mAP
MedFM-Colon Wang et al. (2023)	Pathology	Binary	2	5,654	7,651	Accuracy
MedFM-Endo Wang et al. (2023)	Endoscopy	Multi-label	4	1,810	2,936	mAP

A2 MORE RESULTS

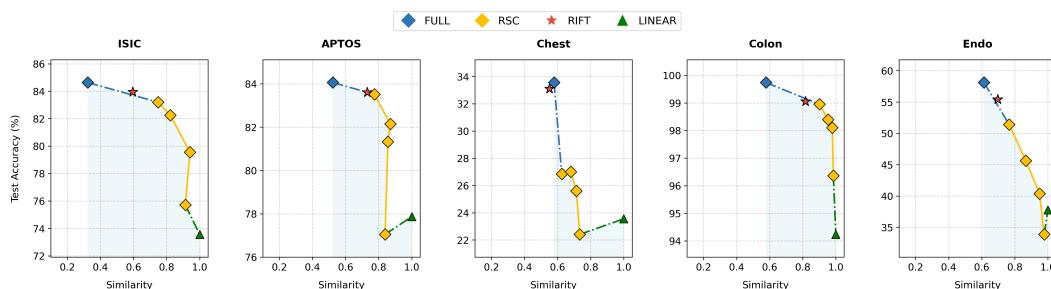


Figure A1: Similarity–accuracy curves of the representation similarity constrained (RSC) method across different datasets. Models above the curve achieve a comparable balance between similarity preservation and accuracy. The λ values range from 4.0 to 0.3, indicating a gradual decrease in the weight of the similarity loss.

Similarity and accuracy curve. As shown in Fig. A1, our proposed RIFT method consistently demonstrates an effective balance between similarity (sim) and accuracy (acc) across multiple datasets. Direct finetuning, the RSC method (λ ranges from 4.0 to 0.3), and finetuning only the linear classification head form the trade-off curve between classification accuracy and representation similarity (with pretrained model). Models above the curve indicate a better balance between accuracy and similarity, while those below the curve show the opposite. Direct finetuning achieves the highest classification accuracy but the lowest representation similarity. Additionally, the trade-off curve

756 Table A2: Detailed evaluation results of RIFT and RSC with different regularization coefficients λ of
 757 Tab. 1. The selected λ corresponds to the parameter yielding the highest accuracy/mAP.
 758

759 coefficient λ	760 Single-label datasets						761 Multi-label datasets			
	762 ISIC2018 (7)		763 APTOS2019 (5)		764 MedFM-Colon (2)		765 MedFM-Chest (19)		766 MedFM-Endo (4)	
767 Acc	768 Sim	769 Acc	770 Sim	771 Acc	772 Sim	773 mAP	774 Sim	775 mAP	776 Sim	
RIFT Method										
$\lambda = 1.0$	83.95	0.60	83.61	0.73	99.05	0.82	33.11	0.55	55.42	0.70
$\lambda = 0.8$	83.73	0.53	84.43	0.76	99.12	0.84	32.22	0.59	55.16	0.70
$\lambda = 0.6$	84.85	0.48	85.25	0.54	99.51	0.84	35.24	0.62	53.66	0.72
$\lambda = 0.4$	84.33	0.50	84.70	0.71	99.51	0.83	34.01	0.53	56.45	0.67
$\lambda = 0.2$	84.57	0.38	83.47	0.53	99.44	0.82	34.47	0.56	57.38	0.67
$\lambda = 0.1$	84.82	0.35	84.84	0.59	99.41	0.83	34.02	0.44	58.62	0.66
RSC Method										
$\lambda = 4.0$	75.71	0.91	77.05	0.84	96.37	0.98	22.41	0.73	33.88	0.98
$\lambda = 1.5$	79.56	0.94	81.33	0.86	98.10	0.98	25.60	0.71	40.38	0.95
$\lambda = 0.7$	82.25	0.82	82.15	0.87	98.40	0.95	27.01	0.68	45.63	0.87
$\lambda = 0.3$	83.20	0.75	83.52	0.77	98.95	0.90	26.86	0.62	51.41	0.77
$\lambda = 0.2$	83.47	0.35	82.79	0.53	99.51	0.82	33.73	0.56	57.52	0.67
$\lambda = 0.1$	85.19	0.35	84.97	0.59	99.12	0.83	34.62	0.44	56.26	0.66
LoRA+RIFT Method										
$\lambda = 1.0$	77.51	0.74	79.24	0.79	24.28	0.63	96.07	0.82	39.24	0.76
$\lambda = 0.2$	78.84	0.60	80.05	0.73	25.51	0.59	96.07	0.81	39.69	0.77
$\lambda = 0.1$	80.32	0.46	80.19	0.67	25.77	0.61	96.37	0.79	38.68	0.78
$\lambda = 0.02$	80.92	0.40	81.01	0.61	24.42	0.73	96.17	0.80	39.88	0.77
Adaptformer+RIFT Method										
$\lambda = 1$	77.65	0.74	80.15	0.85	23.82	0.59	95.87	0.86	38.63	0.75
$\lambda = 0.1$	79.17	0.64	80.60	0.81	24.97	0.64	96.07	0.80	40.55	0.71
$\lambda = 0.02$	79.96	0.56	79.51	0.79	25.24	0.65	97.54	0.81	37.16	0.61
VPT+RIFT Method										
$\lambda = 1$	76.32	0.74	78.69	0.85	22.19	0.59	95.48	0.86	37.70	0.75
$\lambda = 0.1$	77.38	0.68	78.96	0.66	24.27	0.56	94.50	0.74	37.68	0.71
$\lambda = 0.02$	78.31	0.50	78.69	0.57	24.58	0.47	95.38	0.72	40.54	0.67

790
 791 Table A3: Detailed evaluation results of RIFT with different regularization coefficients λ from Tab. 2.
 792 The selected λ corresponds to the parameter yielding the highest accuracy/mAP.
 793

794 coefficient λ	795 Single-label datasets						796 Multi-label datasets			
	797 ISIC2018 (7)		798 APTOS2019 (5)		799 MedFM-Colon (2)		800 MedFM-Chest (19)		801 MedFM-Endo (4)	
802 Acc	803 Sim	804 Acc	805 Sim	806 Acc	807 Sim	808 mAP	809 Sim	810 mAP	811 Sim	
ViT-large(ImageNet-21K)+RIFT Method										
$\lambda = 1.0$	83.66	0.60	82.24	0.89	34.39	0.81	99.12	0.73	58.23	0.66
$\lambda = 0.1$	84.13	0.52	83.06	0.87	38.06	0.67	99.51	0.75	56.40	0.68
ViT-base(Dinov2)+RIFT Method										
$\lambda = 1.0$	69.78	0.40	76.50	0.52	12.83	0.62	92.34	0.68	38.18	0.35
$\lambda = 0.1$	79.63	0.45	68.31	0.28	12.76	0.67	99.71	0.48	40.15	0.36

805
 806
 807 highlights the rapid decline in accuracy as similarity increases (with higher weight on RSC). Our
 808 proposed RIFT method shows comparable results on the ISIC2018, APTOS2019, MedFM-Colon, and
 809 MedFM-Endo datasets compared to the RSC. Although on the MedFM-Chest dataset, we failed to
 810 find a balance in the similarity-accuracy trade-off, these overall results strongly demonstrate that RIFT

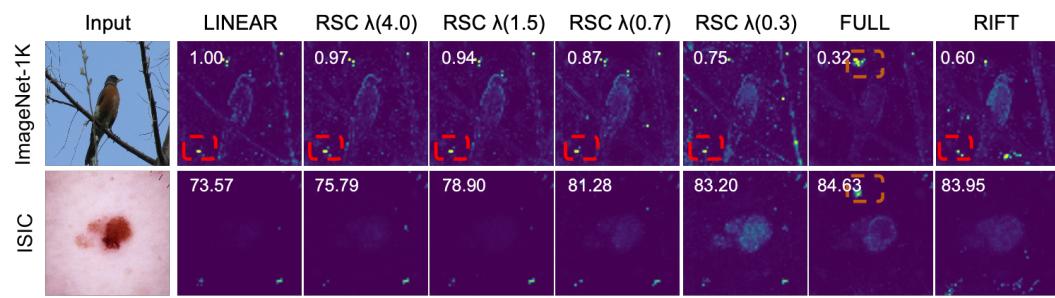


Figure A2: Visualization of attention maps for different finetuning methods on the ISIC2018 dataset. The first row illustrates how the finetuned models forget original knowledge as representation similarity changes, while the second row shows how they learn the new task at the corresponding similarity levels. **This also highlights the potential of RIFT for future applications in continual and multi-task learning.**

Table A4: Comparison of parameter distance to average center for FULL and RIFT methods across five datasets.

Dataset	FULL	RIFT	Change↓
ISIC2018	1.35	0.93	↓ 0.42
APTO2019	0.49	0.45	↓ 0.04
MedFM-Chest	1.02	0.98	↓ 0.04
MedFM-Colon	0.45	0.67	↑ 0.22
MedFM-Endo	0.68	0.58	↓ 0.10
Avg.	0.80	0.72	↓ 0.08

effectively leverages the trade-off between similarity and accuracy, avoiding the severe similarity degradation observed in the direct finetuning method (FULL) when enforcing high accuracy, thereby achieving superior overall performance.

Visualization of attention maps demonstrating learning and forgetting. As shown in Fig. A2, the first row illustrates the changes in the attention maps of the pretrained model as similarity decreases. After finetuning directly on the ISIC2018 dataset (FULL), the attention to the ImageNet-1K image target (bird) vanishes. In contrast, applying a direct similarity constraint (RSC) or an indirect constraint (RIFT) preserves attention to the target object, even retaining artifacts (red dashed box). The second row shows that as classification accuracy improves, the finetuned model gradually enhances its attention to the target object (skin lesion). Notably, the artifacts from the pretrained model are also preserved. For FULL finetuning, the artifacts (orange dashed box) appear in both ImageNet-1K and ISIC2018 images. These observations highlight the necessity of considering similarity constraints.

Enhanced parameter centrality for multi-task learning. Results in Tab. A4 demonstrate that the RIFT method exhibits better parameter centrality across multiple datasets, thereby validating its adaptability in multi-task training. Specifically, the RIFT method significantly reduces the distance of model parameters to the average parameter center on four datasets: ISIC2018, APTOS2019, MedFM-Chest, and MedFM-Endo, with reductions of 0.42, 0.04, 0.04, and 0.10, respectively. Although the distance increases slightly (by 0.22) on the MedFM-Colon dataset, the overall performance of RIFT remains superior to the FULL method, with an average distance reduction of 0.08. These results indicate that the RIFT method can more effectively extract shared features, achieving better parameter centrality and generalization in multi-task training.

Training time comparison. The results in Tab. A5 reveal significant variations in training efficiency across the evaluated methods. The LINEAR method demonstrates the fastest performance, requiring the least time for both per-iteration and full-epoch training. In contrast, the FULL and RIFT methods exhibit nearly identical computational costs, suggesting comparable efficiency. Notably, the RSC

864 Table A5: Training time per iteration and for 50 epochs on single NVIDIA A100 40G with a batch
 865 size of 128.

Metric	LINEAR	FULL	RSC	RIFT
One Iteration (s)	3.91	8.38	16.68	8.47
50 Epochs (min)	26.65	57.08	111.04	57.83

872 method incurs the highest computational overhead, making it the least efficient among the tested
 873 approaches.

874 **Computational Complexity Analysis.** The loss in Eq. equation 11 promotes the transferability of
 875 mean vectors while preserving the structural invariance of covariance matrices under orthogonal
 876 transformations. The orthonormal constraint $\mathbf{Q} \in \mathcal{V}_{d,d}$ guarantees that rotations in the feature space
 877 remain isometric. By decoupling mean and covariance alignment, the computational complexity is
 878 reduced from $O(nd^2)$ to $O(d^2)$ through trace-based operations, while still maintaining the geometric
 879 structure of the representation space. This simplification preserves the essential properties enforced
 880 by orthogonal covariance transformations without introducing additional approximation error.

882 A3 ANALYSIS

884 **Definition A3.1** (Multi-layer linear network with layerwise orthogonal rotations). *Let $X \in \mathbb{R}^{n \times d}$
 885 be the input data with n samples and d features, and let $W_1, W_2, \dots, W_L \in \mathbb{R}^{d \times d}$ be the weight
 886 matrices of an L -layer linear network. Define the network*

$$887 f(X) = XW_1W_2 \cdots W_L \in \mathbb{R}^{n \times d} \quad (\text{A1})$$

888 Consider applying layerwise orthogonal rotations $Q_1, Q_2, \dots, Q_L \in O(d)$ to obtain the rotated
 889 network

$$890 f'(X) = X(W_1Q_1)(W_2Q_2) \cdots (W_LQ_L) = XW' \quad (\text{A2})$$

891 where $O(d)$ denotes the set of $d \times d$ orthogonal matrices. Denote by $\sigma_1(\cdot)$ the largest singular value
 892 (spectral norm) of a matrix.

893 **Assumption A3.2.** Let $W = W_1W_2 \cdots W_L \in \mathbb{R}^{d \times d}$ denote the weight matrix of a pretrained
 894 L -layer linear network. We assume that W exhibits strong generalization, formalized by

$$895 \sigma_1(W) \ll \min_{i=1, \dots, L} \sigma_1(W_i) \quad (\text{A3})$$

896 **Proposition A3.3** (Informal). Let $W = W_1W_2 \cdots W_L$ be an L -layer linear network with $W_i \in$
 897 $\mathbb{R}^{d \times d}$, and assume that W exhibits strong generalization in the sense of Assumption A3.2, i.e.,

$$898 \sigma_1(W) \ll \min_{i=1, \dots, L} \sigma_1(W_i) \quad (\text{A4})$$

900 Then there exist orthogonal matrices $Q_1, \dots, Q_L \in O(d)$ such that the rotated network

$$901 W' = W_1Q_1W_2Q_2 \cdots W_LQ_L \quad (\text{A5})$$

902 has a strictly larger spectral norm than the original W .

903 *Intuitively, because the pretrained network already generalizes well, its layers are not perfectly
 904 aligned to maximize the spectral norm, so an appropriate choice of layerwise rotations can increase
 905 it by improving the alignment of dominant singular directions across layers.*

906 **Assumption A3.4** (Bounded Cross-Covariance). Let X with n samples and feature dimension d ,
 907 and denote the feature matrices $Y = F_\theta(X) \in \mathbb{R}^{n \times d}$, $Y_0 = F_{\theta_0}(X) \in \mathbb{R}^{n \times d}$. Let $Q \in \mathbb{R}^{d \times d}$ be
 908 orthogonal ($Q^\top Q = I_d$) and $\alpha \in \mathbb{R}$, and define the aligned feature matrix $Z := \alpha Y_0 Q$ with residual
 909 $\varepsilon := Y - Z$.

910 For any $A \in \mathbb{R}^{n \times d}$, let $\bar{A} := \frac{1}{n} \mathbf{1}_n^\top A \in \mathbb{R}^{1 \times d}$ denote its column mean, where $\mathbf{1}_n \in \mathbb{R}^n$ is the all-ones
 911 vector, and define the column-centered version $\tilde{A} := A - \mathbf{1}_n \bar{A}$. The empirical cross-covariance
 912 between $A, B \in \mathbb{R}^{n \times d}$ is given by $\text{Cov}(A, B) := \frac{1}{n} \tilde{A}^\top \tilde{B} \in \mathbb{R}^{d \times d}$.

913 By construction, $\text{Cov}(\varepsilon, Y_0) = \text{Cov}(Y, Y_0) - \alpha Q^\top \text{Cov}(Y_0, Y_0)$. We assume this cross-covariance is
 914 bounded as $\|\text{Cov}(\varepsilon, Y_0)\|_F \leq \gamma$ for some constant $\gamma \geq 0$.

918 **Theorem A3.5.** Define $Y := F_\theta(X) \in \mathbb{R}^{n \times d}$, $Y_0 := F_{\theta_0}(X) \in \mathbb{R}^{n \times d}$ as feature matrices of n
 919 samples with d -dimensional features. Denote the row-wise means $\mu_\theta := \frac{1}{n} \mathbf{1}_n^\top Y$, $\mu_{\theta_0} := \frac{1}{n} \mathbf{1}_n^\top Y_0$, and
 920 the empirical covariances $\Sigma_\theta := \frac{1}{n} (Y - \mathbf{1}_n \mu_\theta)^\top (Y - \mathbf{1}_n \mu_\theta)$, $\Sigma_{\theta_0} := \frac{1}{n} (Y_0 - \mathbf{1}_n \mu_{\theta_0})^\top (Y_0 - \mathbf{1}_n \mu_{\theta_0})$.
 921 Let $Q \in \mathbb{R}^{d \times d}$ be orthogonal ($Q^\top Q = I_d$) and $\alpha \in \mathbb{R}$. Define $Z := \alpha Y_0 Q$, $\varepsilon := Y - Z$. Assume the
 922 cross-covariance $\text{Cov}(\varepsilon, Y_0) := \frac{1}{n} (\varepsilon - \mathbf{1}_n \bar{\varepsilon})^\top (Y_0 - \mathbf{1}_n \bar{Y}_0)$ satisfies $\|\text{Cov}(\varepsilon, Y_0)\|_F \leq \gamma$, where $\bar{\varepsilon} :=$
 923 $\frac{1}{n} \mathbf{1}_n^\top \varepsilon$, $\bar{Y}_0 := \frac{1}{n} \mathbf{1}_n^\top Y_0$. Define $\mathcal{E} := \frac{1}{n} \|Y - \alpha Y_0 Q\|_F^2$, $\Delta\mu := \mu_\theta - \alpha \mu_{\theta_0} Q$, $\Delta\Sigma := \Sigma_\theta - \alpha^2 Q^\top \Sigma_{\theta_0} Q$.
 924 Then

$$|\mathcal{E} - (\|\Delta\mu\|_2^2 + \text{tr}(\Delta\Sigma))| \leq 2|\alpha|\sqrt{d}\gamma \quad (\text{A6})$$

927 *Proof.* Let

$$\varepsilon = Y - Z = Y - \alpha Y_0 Q \quad (\text{A7})$$

928 Its row-wise mean is

$$\bar{\varepsilon} = \frac{1}{n} \mathbf{1}_n^\top \varepsilon = \mu_\theta - \alpha \mu_{\theta_0} Q = \Delta\mu \quad (\text{A8})$$

929 Define the centered matrices

$$\tilde{Y} = Y - \mathbf{1}_n \mu_\theta, \quad \tilde{Y}_0 = Y_0 - \mathbf{1}_n \mu_{\theta_0}, \quad \tilde{\varepsilon} = \varepsilon - \mathbf{1}_n \bar{\varepsilon} = \tilde{Y} - \alpha \tilde{Y}_0 Q \quad (\text{A9})$$

930 so that

$$\text{Cov}(\varepsilon) = \frac{1}{n} \tilde{\varepsilon}^\top \tilde{\varepsilon} \quad (\text{A10})$$

931 Using $Y = Z + \varepsilon$ and bilinearity of covariance, we have

$$\text{Cov}(Y) = \text{Cov}(Z) + \text{Cov}(\varepsilon) + \text{Cov}(\varepsilon, Z) + \text{Cov}(Z, \varepsilon) \quad (\text{A11})$$

932 Since $\tilde{Z} = \alpha \tilde{Y}_0 Q$, it follows that

$$\text{Cov}(Z) = \frac{1}{n} \tilde{Z}^\top \tilde{Z} = \alpha^2 Q^\top \Sigma_{\theta_0} Q \quad (\text{A12})$$

933 Therefore

$$\Delta\Sigma = \text{Cov}(Y) - \text{Cov}(Z) = \text{Cov}(\varepsilon) + \text{Cov}(\varepsilon, Z) + \text{Cov}(Z, \varepsilon) \quad (\text{A13})$$

934 and hence

$$\text{tr}(\Delta\Sigma) = \text{tr}(\text{Cov}(\varepsilon)) + 2 \text{tr}(\text{Cov}(\varepsilon, Z)) \quad (\text{A14})$$

935 On the other hand, the error can be written as

$$\mathcal{E} = \frac{1}{n} \|Y - \alpha Y_0 Q\|_F^2 = \frac{1}{n} \|\varepsilon\|_F^2 = \|\Delta\mu\|_2^2 + \text{tr}(\text{Cov}(\varepsilon)) \quad (\text{A15})$$

936 Subtracting $\|\Delta\mu\|_2^2 + \text{tr}(\Delta\Sigma)$ gives

$$\mathcal{E} - (\|\Delta\mu\|_2^2 + \text{tr}(\Delta\Sigma)) = -2 \text{tr}(\text{Cov}(\varepsilon, Z)) \quad (\text{A16})$$

937 Now observe that

$$\text{Cov}(\varepsilon, Z) = \frac{1}{n} \tilde{\varepsilon}^\top \tilde{Z} = \frac{1}{n} \tilde{\varepsilon}^\top (\alpha \tilde{Y}_0 Q) = \alpha \text{Cov}(\varepsilon, Y_0) Q \quad (\text{A17})$$

938 Therefore

$$|\text{tr}(\text{Cov}(\varepsilon, Z))| = |\alpha| |\text{tr}(\text{Cov}(\varepsilon, Y_0) Q)| \leq |\alpha| \sqrt{d} \|\text{Cov}(\varepsilon, Y_0)\|_F \leq |\alpha| \sqrt{d} \gamma \quad (\text{A18})$$

939 where we used $|\text{tr}(AQ)| \leq \|A\|_* \leq \sqrt{d} \|A\|_F$

940 Combining the results, we conclude

$$|\mathcal{E} - (\|\Delta\mu\|_2^2 + \text{tr}(\Delta\Sigma))| = 2 |\text{tr}(\text{Cov}(\varepsilon, Z))| \leq 2|\alpha|\sqrt{d}\gamma \quad (\text{A19})$$

941 \square

942 **Remark A3.6** (Interpretation and Significance). *Theorem A3.5 shows that the RIFT objective, expressed in terms of the mean difference $\Delta\mu$ and the covariance difference $\Delta\Sigma$, accurately approximates the original alignment loss \mathcal{E} . Specifically, if the residual ε has zero cross-covariance with the pretrained representation Y_0 (i.e., $\gamma = 0$), the RIFT objective equals the true alignment loss exactly. When small cross-covariances exist, the discrepancy is explicitly bounded by $2|\alpha|\sqrt{d}\gamma$, depending only on the feature dimension and the cross-covariance magnitude. This provides theoretical support for using covariance-based orthogonal alignment: it eliminates the need to compute \mathcal{E} over all samples while retaining the key alignment characteristics of the original representation space.*