Selective Rank-1 Orthogonality Regularization: Rethinking Stability-Plasticity Balance in Continual Medical Image Classification

Jingyang Zhang[1], Yiqing Guo[1], Jiarui Sun[2], Dunyuan Xu[3], Shuo Gao[4(✉)], and Lei Ma[5(✉)]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China
[2] School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing, China
[3] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
[4] School of Biological Science and Medical Engineering, Southeast University, Nanjing, China
shuogao@seu.edu.cn,
[5] School of Electronics and Information Engineering, Tongji University, Shanghai, China
ma_lei@tongji.edu.cn

Abstract. Faced with the emergence of new diseases in clinical practice, Continual Learning (CL) requires models to learn tasks sequentially with a critical balance between plasticity (new-task adaptation) and stability (prior-task retention). Pre-Trained Models (PTMs) offer a generalizable foundation for CL with compact parameter spaces injected incrementally, but limited by a stability-plasticity dilemma due to their uniform constraint on the whole space, e.g., full orthogonality or task-specific update. Hence, we introduce a principled subspace decomposition perspective to fundamentally revisit this dilemma, motivating a novel Selective Rank-1 Orthogonality Regularization (SR1OR) framework for stability-plasticity balance. Specifically, by exhaustively decomposing LoRA into unit rank-1 subspaces, we establish a Rank-1 Subspace Orientation Theory (R1SOT) that justifies unique orientation of each subspace based on subspace-local curvature and task-global sharpness. Guided by this, we develop Progressive Subspace Identification (PSI) to identify stability-oriented subspaces with large local curvature especially for tasks with high global sharpness. When new tasks arrive, we impose a Selective Orthogonality Constraint (SOC) only on these subspaces to mitigate interference for high stability, without regularization on residual subspaces to accommodate adaptation plasticity. Experiments on three class-incremental diagnosis benchmarks show a clear advantage for stability-plasticity balance than SOTA methods. Code is available at https://github.com/jingyzhang/SR1OR/.

Keywords: Continual Learning, Stability, Plasticity, Orthogonality

1 Introduction

In clinical practice, deep learning-based classification methods struggle with the emergence of new diseases as diagnosis tasks evolve [15,18]. Moreover, data from prior tasks is often inaccessible due to privacy and storage limitation [16], making these models prone to model forgetting [25] with high performance drops on prior tasks. Therefore, Continual Learning (CL) is developed to cultivate a generalist diagnostic skill across sequential tasks [12], pursuing a delicate balance between plasticity (adapting to new tasks) and stability (retaining prior-task knowledge).

Recently, large-scaled Pre-Trained Models (PTMs) have served as a starting point for CL [23] through freezing them as backbones to retain general knowledge [8], while parameter-efficient finetuning methods (e.g., LoRA [10]) consecutively adapt them to new tasks by incrementally adding parameter spaces in two ways: 1) Task-specific optimization (as shown in Fig. 1(a)) allocates separate parameter spaces for each new task [23] and finetunes them individually for high adaptation plasticity. However, it can lead to forgetting of prior knowledge in the overlapping space and weaken stability for knowledge accumulation, which especially relies on accurate task IDs for space retrieval at inference [22,19]; and 2) Full orthogonality constraint (as shown in Fig. 1(b)) expands strictly orthogonal parameter spaces across tasks without overlapping to strengthen stability [21,13], but under a fixed dimensionality, it leaves limited residual capacity and reduces new-task plasticity. Overall, these PTM-based CL methods suffer from a stability-plasticity dilemma rooted in *a uniform perspective on the whole parameter space*: full orthogonality rapidly occupies the space that limits plasticity, while task-specific optimization hinders stability because each entire space is updated free of regularization.

To solve the dilemma, our insight is to impose *selective orthogonality over unit parameter subspaces for each task*, as shown in Fig. 1(c), which provides a flexible alternative to the space-uniform perspective that leads to either overly-strict full orthogonality or overly-loose unconstrained task-specific optimization. This idea is motivated by PTM scaling laws [2,20], suggesting that a cellular subspace decomposition of the parameter space can make the subspace count a measurable proxy for model capacity, spanning from stable retention to plastic adaption [14]. But, the subspace-set size acts only as a coarse capacity proxy and even cannot specify which subspaces to select for a fine-grained stability-plasticity balance in CL, due to the lack of subspace-specific analysis on unique orientation regarding stability or plasticity. Thus, our goal is to establish a theoretical guarantee that each unit subspace admits an intrinsic orientation that could be characterized individually. This theory, in turn, justifies enforcing selective orthogonality only on stability-oriented subspaces to avoid knowledge overwriting, while dynamically releasing the remaining plasticity-oriented subspaces for new-task adaptation.

In this work, we present *to our knowledge the first theoretical viewpoint* on the stability-plasticity dilemma for PTM-based CL, rooted in a parameter-subspace decomposition perspective. Specifically, building upon the widely-used LoRA for low-rank adaptation over tasks, we propose a novel Selective Rank-1 Orthogonality Regularization (SR1OR) framework for CL, which makes three contributions: 1) By decomposing each task's LoRA space into unit rank-1 subspaces, we build a

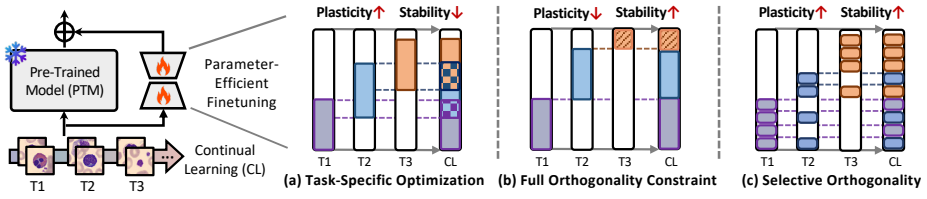

Fig. 1. Illustration of PTM-based CL with parameter spaces added incrementally via: a) Task-specific optimization accumulates entire spaces over time but causes knowledge overlap; b) Full orthogonality is constrained on the complete space across tasks without enough spaces residual for new-task adaptation; and c) Our selective orthogonality uses a cellular subspace decomposition and enforces orthogonality only on partial subspaces oriented to stability, while leaving the remaining subspaces unconstrained for plasticity.

Rank-1 Subspace Orientation Theory (R1SOT) to justify an individual stability- or plasticity-oriented role for each subspace, by characterizing the loss landscape in terms of local curvature in each subspace and global sharpness for a given task; 2) Guided by this theory, we develop a Progressive Subspace Identification (PSI) method to distinguish stability-oriented subspaces via a local-to-global scheme, prioritizing high-curvature subspaces selected in tasks with sharp loss landscape; and 3) Given subspaces identified for prior tasks, we design a Selective Orthogonality Constraint (SOC) to regularize new-task adaptation with orthogonality to stability-oriented subspaces to minimize interference and forgetting, while leaving plasticity-oriented subspaces free to update for high model capacity. We have evaluated our SR1OR on three class-incremental diagnosis benchmarks, showing its better plasticity-stability balance than the exiting PTM-based CL methods.

2 Method

Given a sequence of tasks with new classes arriving over time, the model learns the incoming task $\mathcal{D}_{k+1}$ incrementally, without access to the prior tasks $\{\mathcal{D}_i\}_{i=1}^{k}$. Fig. 2 illustrates our Selective Rank-1 Orthogonality Regularization (SR1OR), which decomposes and identifies stability-oriented subspaces for prior tasks to derive a selective orthogonality constraint without sacrificing new-task plasticity.

2.1 Preliminary on Low-Rank Adaptation (LoRA) for CL

Considering the pre-trained backbone as a generalizable starting point, parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) [10] offers a practical way for CL. Let $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ denote the frozen weight in a backbone layer. When task $\mathcal{D}_i$ arrives, we expand a new LoRA branch that parameterizes a low-rank update with learnable matrices $\mathbf{A}_i \in \mathbb{R}^{m\times r}$ and $\mathbf{B}_i \in \mathbb{R}^{r\times n}$, where rank $r \ll \min\{m, n\}$. After CL across tasks $\{\mathcal{D}_i\}_{i=1}^{k}$, the weight $\mathbf{W}_k$ aggregates these low-rank updates and integrates as a single branch for inference without task-specific selection [24]:

$$\mathbf{W}_k = \mathbf{W}_{k-1} + \mathbf{A}_k\mathbf{B}_k = \mathbf{W}_0 + \sum\nolimits_{i=1}^{k} \mathbf{A}_i\mathbf{B}_i. \tag{1}$$

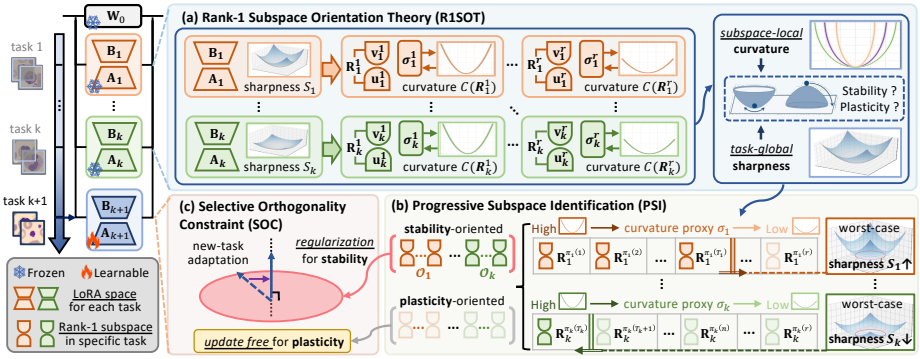

Fig. 2. Diagram of our Selective Rank-1 Orthogonality Regularization (SR1OR) framework, building on a naive LoRA-based CL baseline (Sect. 2.1). Concretely, based on unit subspace decomposition for prior-task's LoRA, we establish a Rank-1 Subspace Orientation Theory (R1SOT) to warrant their intrinsic roles in CL (Sect. 2.2), and adopt Progressive Subspace Identification (PSI) to distinguish stability- versus plasticity-oriented subspaces from prior tasks under this theory (Sect. 2.3). When a new task arrives, only stability-oriented subspaces are involved in a Selective Orthogonality Constraint (SOC) to prevent forgetting, while reserving the remaining subspaces for plasticity (Sect. 2.4).

2.2 Rank-1 Subspace Orientation Theory (R1SOT)

While LoRA offers a feasible CL capacity, cross-task updates can become entangled in the unified low-rank space, making it hard to preserve prior-task stability without sacrificing new-task plasticity. This problem stems from the lack of fine-grained exploration of the unified low-rank parameter space, which inspires us to exhaustively decompose it into unit rank-1 subspaces and delve into their intrinsic orientations, laying a theoretical foundation for plasticity-stability trade-off.

Unit Rank-1 Subspace Decomposition. For task $\mathcal{D}_i$ with rank-r space $\mathbf{A}_i\mathbf{B}_i$, we decompose it into r unit rank-1 subspace $\{\mathbf{R}_i^j\}_{j=1}^r$ based on the Singular Value Decomposition (SVD). Specifically, by rewriting singular vector matrices $\mathbf{U}_i = [\mathbf{u}_i^1, ..., \mathbf{u}_i^r]$ and $\mathbf{V}_i^T = [\mathbf{v}_i^1, ..., \mathbf{v}_i^r]^T$, each rank-1 subspace is given by $\mathbf{R}_i^j = \mathbf{u}_i^j(\mathbf{v}_i^j)^T$ and weighted by its singular value σ_i^j in the matrix $\mathbf{\Sigma}_i = \text{diag}(\sigma_i^1, ..., \sigma_i^r)$:

$$\mathbf{A}_i\mathbf{B}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^T = \sum\nolimits_{j=1}^{r} \sigma_i^j \mathbf{R}_i^j, \quad \text{where} \quad \mathbf{R}_i^j = \mathbf{u}_i^j(\mathbf{v}_i^j)^T. \tag{2}$$

The subspaces are strictly disjoint by SVD orthonormality, avoiding redundancy in the original row-column space [20] and grounding independent analysis below.

Theoretical Analysis on Subspace Orientation. To probe the characteristic of each rank-1 subspace, we formulate a subspace shift bound theorem based on Empirical Risk Minimization (ERM) as tasks arrive sequentially. Heuristically, a large subspace shift indicates high plasticity that requires more updates for new-task adaptation, while a small shift suggests stability with minimal modification.

Theorem 1 (Subspace Shift Bound). *Consider a model* $\mathbf{W}_k$ *after CL across tasks* $\{\mathcal{D}_i\}_{i=1}^k$ *with rank-1 subspaces* $\{\mathbf{R}_i^j\}_{j=1}^r$ *decomposed for each prior task. Assume the new-task adaptation is restricted to shifting each subspace by a learnable scale factor* α_i^j*, formulated as* $\mathbf{\Delta}_\alpha^{\mathbf{R}}\mathbf{W}_k = \mathbf{W}_k + \sum_{i=1}^k \sum_{j=1}^r \alpha_i^j \mathbf{R}_i^j$*.* α_i^j *is learned by minimizing the new-task loss* $\mathcal{L}_{k+1}(\cdot)$ *and penalizing increases in prior-task losses* $\{\mathcal{L}_i(\cdot)\}_{i=1}^k$*, each weighted by its sharpness* S_i *to suppress rapid variations [5]:*

$$\underset{\{\alpha\}}{argmin}\ \mathcal{L}_{k+1}(\mathbf{\Delta}_\alpha^{\mathbf{R}}\mathbf{W}_k) + \sum\nolimits_{i=1}^k S_i \left[\mathcal{L}_i(\mathbf{\Delta}_\alpha^{\mathbf{R}}\mathbf{W}_k) - \mathcal{L}_i(\mathbf{W}_k)\right]. \tag{3}$$

Based on a second-order Taylor expansion with a Lagrangian multiplier augmentation, the subspace shift scale α_i^j *admits a closed-form bound, given a gradient-clip constant* κ*:*

$$|\alpha_i^j| \leq \frac{\kappa}{C(\mathbf{R}_i^j) \cdot S_i}, \tag{4}$$

where $C(\mathbf{R}_i^j)$ *denotes the local curvature of the loss landscape along subspace* $\mathbf{R}_i^j$*, and* S_i *denotes the global sharpness of the loss landscape specific to task* $\mathcal{D}_i$*.*

Intuitively, for subspace $\mathbf{R}_i^j$, its loss-landscape geometry with high subspace-local curvature $C(\mathbf{R}_i^j)$ and task-global sharpness S_i can lead to a narrow shift scale α_i^j, indicating a stability orientation with minor modification required. Conversely, lower $C(\mathbf{R}_i^j)$ and S_i can relax the shift bound to permit large new-task plasticity.

Rethinking Plasticity-Stability Dilemma. In light of subspace decomposition, existing PTM-based CL often adopts a subspace-uniform treatment. Concretely, full orthogonality constraints are imposed on most (or even all) subspaces to ensure perfect stability [21,13], at the expense of reserving limited capacity for plasticity. Moreover, task-specific optimization assumes uniformly high plasticity for each subspace by allowing diverse new-task shifts [22], instead of consolidating for stability. Overall, they overlook intrinsic orientations across subspaces, which should inherently vary with subspace-local curvature and task-global sharpness.

2.3 Progressive Subspace Identification (PSI)

Instead of imposing a subspace-uniform viewpoint, we identify each rank-1 subspace by its unique orientation for stability-plasticity balance. Naively, this can be regarded as a dual-factor ranking problem over subspace-local curvature and task-global sharpness, which is hard to solve via Pareto parallel optimization [4]. Hence, we propose a Progressive Subspace Identification (PSI) method in a local-to-global way, which ranks subspaces in each task by descending singular value as a curvature proxy, and then selects top-curvature subspaces as stability-oriented using an adaptive threshold that even favors tasks with sharper loss landscapes.

Within-Task Ranking. Stability-oriented subspaces are expected to have high local curvature in the loss landscape for a given task. Since directly computing curvature is costly, we utilize each subspace's singular value in Eq. (2) as an efficient proxy $\sigma_i^j \propto C(\mathbf{R}_i^j)$, capturing the principal curvature geometry by Hessian

spectrum analysis [9]. Formally, in each task $\mathcal{D}_i$, we rank subspaces by singular value in descent order $\mathit{\Pi}_i$, where higher-ranked subspaces have higher curvature:

$$\mathit{\Pi}_i = \{\pi_i(j)\}_{j=1}^{r}, \quad \text{where} \quad \sigma_i^{\pi_i(1)} \geq \sigma_i^{\pi_i(2)} \geq \cdots \geq \sigma_i^{\pi_i(r)}. \tag{5}$$

Task-Adaptive Thresholding. Beyond subspace-local curvature, subspace orientation also depends on the task in which the subspace resides, as characterized by the task-global sharpness of the loss landscape. In principle, stability-oriented subspaces should be not only top-ranked for high curvature, but also more prevalent in sharper tasks. Accordingly, for task $\mathcal{D}_i$ with subspaces ranked by $\mathit{\Pi}_i$, we normalize its sharpness to obtain $\bar{S}_i \in [0, 1]$, and use it to set an adaptive threshold $\mu_i(\bar{S}_i) = \mu_{\min} + (\mu_{\max} - \mu_{\min})\bar{S}_i$ that adjusts the subspace-selection ratio in a predefined budget $[\mu_{\min}, \mu_{\max}]$. Based on cumulative energy of singular values [17], we select stability-oriented subspaces $\mathcal{O}_i$ from plasticity-oriented residuals:

$$\mathcal{O}_i = \left\{\mathbf{R}_i^{\pi_i(j)}\right\}_{j=1}^{T_i}, \quad \text{with} \quad T_i = \operatorname*{argmin}_t \left\{ \frac{1}{Z_i} \sum\nolimits_{j=1}^{t} \left(\sigma_i^{\pi_i(j)}\right)^2 \leq \mu_i(\bar{S}_i) \right\}, \tag{6}$$

where Z_i is the total energy accumulated over all r subspaces in task $\mathcal{D}_i$. Notably, we estimate the sharpness $S_i = \max_{\|\boldsymbol{\epsilon}\|_F \leq \rho} \{\mathcal{L}_i(\mathbf{W}_i + \boldsymbol{\epsilon}) - \mathcal{L}_i(\mathbf{W}_i)\}$ via the worst-case degradation in a perturbation radius ρ [7], which can be computed efficiently after each prior CL round with the learned parameter $\mathbf{W}_i$ and its loss $\mathcal{L}_i(\cdot)$. With this design, $\mathcal{O}_i$ contains stability-oriented subspaces that satisfy both factors in R1SOT: they exhibit high local curvature (top-ranked by $\sigma_i^{\pi_i(j)}$) and are selected more aggressively for tasks with sharper loss landscapes (larger S_i).

2.4 Selective Orthogonality Constraint (SOC)

When learning on a new task $\mathcal{D}_{k+1}$, we expand a new LoRA branch $\mathbf{A}_{k+1}\mathbf{B}_{k+1}$ to the previous parameter $\mathbf{W}_k$ by optimize the new-task loss $\mathcal{L}_{k+1}(\cdot)$. For stability-plasticity balance, we propose a Selective Orthogonality Constraint (SOC) $\mathcal{L}_{soc}(\cdot)$ that enforces the new-task adaptation $\mathbf{A}_{k+1}\mathbf{B}_{k+1}$ to be orthogonal to the stability-oriented subspaces $\mathbf{R}_i^j \in \mathcal{O}_i$ identified for each prior task to suppress shifts, while keeping unconstrained capacity in the residual subspaces for new-task plasticity.

$$\operatorname*{argmin}_{\mathbf{A}_{k+1}, \mathbf{B}_{k+1}} \mathcal{L}_{k+1}(\mathbf{W}_k + \mathbf{A}_{k+1}\mathbf{B}_{k+1}) + \lambda \sum\nolimits_{i=1}^{k} \sum\nolimits_{\mathbf{R}_i^j \in \mathcal{O}_i} \mathcal{L}_{soc}\left(\mathbf{R}_i^j, \mathbf{A}_{k+1}\mathbf{B}_{k+1}\right) \tag{7}$$

where λ controls the strength of $\mathcal{L}_{soc}$. Moreover, directly imposing orthogonality between $\mathbf{R}_i^j$ and $\mathbf{A}_{k+1}\mathbf{B}_{k+1}$ is numerically unstable due to high dimensionality. Noting that updates in $\mathbf{A}_{k+1}\mathbf{B}_{k+1}$ can be spanned by the column space of $\mathbf{A}_{k+1}$ (with $\mathbf{B}_{k+1}$ as coefficients) [21], we simplify $\mathcal{L}_{soc}$ in the column space $\mathbf{A}_{k+1}$ with respect to $\mathbf{u}_i^j$ for the subspace $\mathbf{R}_i^j = \mathbf{u}_i^j (\mathbf{v}_i^j)^T$, using inner-product with L2-norm:

$$\mathcal{L}_{soc}\left(\mathbf{R}_i^j, \mathbf{A}_{k+1}\mathbf{B}_{k+1}\right) = \|(\mathbf{u}_i^j)^T \mathbf{A}_{k+1}\|_2. \tag{8}$$

3 Experiments

Dataset. We evaluated our method on three class-incremental diagnosis datasets: 1) peripheral blood cell dataset [1] contains 17092 microscopic images of 8 cell classes, divided into 4 tasks for CL with 2 classes added at each task; 2) colon tissue dataset [11] includes 107180 histological images for 9 tissue types, setting up 4 CL tasks with class increments of [3,2,2,2]; and 3) fundus disease dataset [3] has 1000 fundus images from 39 disease classes, divided into 4 CL tasks with [8,8,8,7] classes involved. All images were resized to 224×224 and unit-normalized. These datasets were split into 70%, 10% and 20% for training, validation and testing.

Evaluation Metrics. After training on the final task, we report the overall CL performance using Averaged accuracy (AVG) over both prior and incoming tasks. In addition, we assess the stability-plasticity balance using two standard metrics [26]. Backward Transfer (BWT) quantifies stability by calculating the drop in accuracy on earlier tasks after learning subsequent tasks. Transfer Learning (TL) evaluates plasticity by aggregating accuracy on newly-involved tasks throughout CL. A strong method is expected to achieve high AVG, high TL and low BWT.

Implementation Details. We adopted a pre-trained ViT-B/16 [6] as the backbone. As suggested by [8], LoRA was injected into the query and value projection in multi-head attention blocks with a fixed rank $r = 10$. After CL on prior tasks, we identified stability-oriented subspaces in PSI using an adaptive threshold in a budget $[0.5, 0.8]$ with loss sharpness estimated in a radius $\rho = 0.5$. When new tasks arrive, we set $\lambda = 10$ to control SOC weight, and expand a new LoRA branch to update the model, which can be directly used for inference over classes learned so far. The model was trained by the Adam Optimizer for 20 epochs with learning rate $1{\times}10^{-3}$ and batch size 16.

Comparison with State-of-the-Arts. We compared our SR1OR method with recent PTM-based CL methods: 1) *Baseline*: **FineTune** on sequentially-arriving tasks without applying CL schemes; 2) *Task-specific optimization schemes*: **L2P** [23] optimizing prompt spaces allocated to each task, and **DualPrompt** [22] and **CODA-Prompt** [19] consolidating these spaces by prompt disentanglement and consistency collaboration; and 3) *Full orthogonality constraint schemes*: **OLoRA** [21] and **InfLoRA** [13] imposing the strict orthogonality regularization throughout the entire parameter space and feature embedding, and **SD-LoRA** [24] using knowledge distillation to guarantee orthogonal update directions.

As listed in Table 1, FineTune shows poor performance with the lowest AVG, as its stability is severely eroded for the largest BWT. Task-specific optimization improves stability yet at the cost of plasticity with lower TL, which is exacerbated for full orthogonality constraints that rapidly sacrifice TL despite decent BWT. Our SR1OR gains the highest AVG for the best trade-off, e.g., for the blood cell, it outperforms CODA-Prompt by 7.59% BWT and InfLoRA by 12.59% TL.

Ablation Study. As shown in Fig. 3(a), removing all modules degenerates our method to FineTune with a poor performance. When adding SOC for CL regularization, model forgetting can be relieved for lower BWT but plasticity inevitably

Table 1. Performance comparison on datasets for blood cell, colon tissue and fundus disease. Metrics are based on accuracy (%), with the best results highlighted in **Bold**.

		Blood Cell			Colon Tissue			Fundus Disease		
		AVG ↑	TL ↑	BWT ↓	AVG ↑	TL ↑	BWT ↓	AVG ↑	TL ↑	BWT ↓
Baseline	FineTune	46.89	93.20	51.27	62.89	86.80	25.81	36.69	68.49	39.10
Task-Specific Optimization	L2P [23]	62.71	85.13	23.27	78.87	86.40	9.50	40.55	61.55	27.53
	DualPrompt [22]	60.57	88.83	29.33	74.90	86.54	12.67	44.71	61.93	21.23
	CODA-Prompt [19]	64.41	89.04	25.52	77.71	86.63	9.94	44.11	57.42	30.80
Full Orthogonality Constraint	OLoRA [21]	65.17	82.41	19.18	78.37	81.89	**5.02**	43.41	53.91	20.45
	InfLoRA [13]	68.72	81.78	18.65	70.72	77.45	8.13	47.55	55.66	20.41
	SD-LoRA [24]	66.93	79.17	21.40	76.84	79.64	5.71	45.51	55.83	17.39
Proposed	SR1OR (ours)	**76.94**	**94.37**	**17.93**	**83.12**	**87.06**	5.13	**56.07**	**68.94**	**16.32**

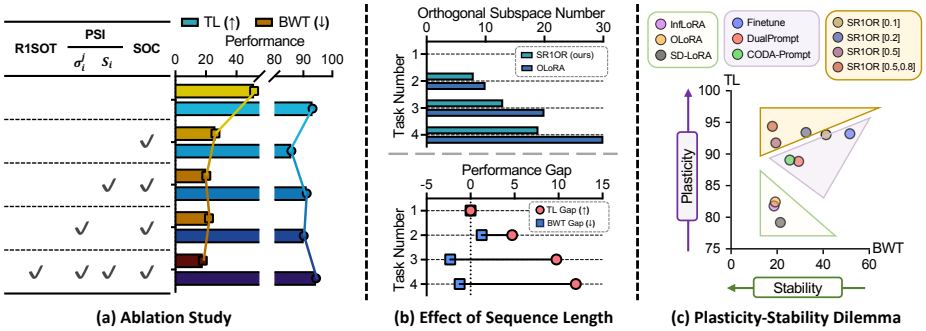

Fig. 3. Model analysis on the blood cell dataset: (a) Ablation study on each module; (b) Effect of sequence length on the orthogonal subspace numbers in SR1OR and OLoRA and their resulting performance gaps; and (c) Balance between plasticity and stability.

drops with limited TL, since orthogonal subspaces are randomly selected without orientation analysis. Moreover, utilizing PSI with either subspace-local curvature (indicated by σ_i^j) or task-global sharpness (measured by S_i) can further improve stability with less BWT and avoid impairing plasticity with simultaneously high TL. Finally, incorporating both factors into PSI can lead to the best performance since it aligns well with the theoretical guidance of R1SOT.

Effect of Learning Sequence Length. Fig. 3(b) indicates how the number of orthogonal subspaces in SR1OR and OLoRA grows as a task sequence increases, and how their performance gap evolves as more tasks are involved. Our SR1OR can expand orthogonal subspaces more slowly than OLoRA owing to our selective constraint. It better reserves plasticity even with more tasks added, as evidenced by the increased TL gap, while BWT keeps stable without aggravating forgetting.

Plasticity-Stability Balance. We compared our SR1OR with other CL methods in Fig. 3(c), evaluating plasticity and stability by TL and BWT, respectively. Full orthogonality constraints (e.g., InfLoRA) favor stability yet limit plasticity (located in the lower left), whereas task-specific optimization (e.g., DualPrompt) boosts plasticity at the cost of increased forgetting (located in the upper right). Our SR1OR lies in the upper-left region with a better plasticity-stability balance,

where an adaptive threshold in [0.5, 0.8] outperforms fixed low ones (especially for 0.1) that select too few stability-oriented subspaces and collapse to FineTune.

4 Conclusion

This paper demonstrates the plasticity-stability dilemma of existing PTM-based CL methods rooted in their strict space-uniform perspective. Hence, we provide a flexible subspace decomposition viewpoint and propose a novel Selective Rank-1 Orthogonality Regularization (SR1OR) framework, which identifies stability-oriented subspaces for selective orthogonality, following a theoretical foundation for plasticity-stability balance. Our broader impact is to build a generalist diagnostic platform across sequential tasks while keeping plasticity for rapid update. In the future, we will investigate its application to domain-incremental scenarios, e.g., sequential medical sites or modalities, and its scalability for more backbones.

Acknowledgments. This work is funded by the National Natural Science Foundation of China (62501146, 62303320), the Basic Research Program of Jiangsu (BK20251345), Research Start Up Grant of Southeast University (4009002412), and the Civil Space Technology Pre-research Foundation(D010101).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in Brief **30**, 105474 (2020)
2. Aghajanyan, A., Gupta, S., Zettlemoyer, L.: Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). pp. 7319–7328 (2021)
3. Cen, L.P., Ji, J., Lin, J.W., Ju, S.T., Lin, H.J., Li, T.P., Wang, Y., Yang, J.F., Liu, Y.F., Tan, S., et al.: Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature Communications **12**(1), 4828 (2021)
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A fast and elitist multi-objective genetic algorithm: Nsga-ii. IEEE Transactions on Evolutionary Computation **6**(2), 182–197 (2002)
5. Deng, D., Chen, G., Hao, J., Wang, Q., Heng, P.A.: Flattening sharpness for dynamic gradient projection memory benefits continual learning. Advances in Neural Information Processing Systems **34**, 18710–18721 (2021)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations (2021)

7. Foret, P., Kleiner, A., Mobahi, H., Neyshabur, B.: Sharpness-aware minimization for efficiently improving generalization. International Conference on Learning Representations (2021)
8. Gao, Q., Zhao, C., Sun, Y., Xi, T., Zhang, G., Ghanem, B., Zhang, J.: A unified continual learning framework with general parameter-efficient tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11483–11493 (2023)
9. Ghorbani, B., Krishnan, S., Xiao, Y.: An investigation into neural net optimization via hessian eigenvalue density. In: International Conference on Machine Learning. pp. 2232–2241. PMLR (2019)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. International Conference on Learning Representations (2022)
11. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Medicine **16**(1), e1002730 (2019)
12. Li, W., Zhang, J., Heng, P.A., Gu, L.: Comprehensive generative replay for task-incremental segmentation with concurrent appearance and semantic forgetting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 80–90. Springer (2024)
13. Liang, Y.S., Li, W.J.: Inflora: Interference-free low-rank adaptation for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23638–23647 (2024)
14. Lu, H., Zhao, C., Xue, J., Yao, L., Moore, K., Gong, D.: Adaptive rank, reduced forgetting: Knowledge retention in continual learning vision-language models with dynamic rank-selective lora. arXiv preprint arXiv:2412.01004 (2024)
15. Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G.: Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature Communications **12**(1), 5678 (2021)
16. Price, W.N., Cohen, I.G.: Privacy in the age of medical big data. Nature Medicine **25**(1), 37–43 (2019)
17. Qiu, H., Zhang, M., Qiao, Z., Guan, W., Zhang, M., Nie, L.: Splitlora: Balancing stability and plasticity in continual learning through gradient space splitting. arXiv preprint arXiv:2505.22370 (2025)
18. Rajpurkar, P., Chen, E., Banerjee, O., Topol, E.J.: AI in health and medicine. Nature Medicine **28**(1), 31–38 (2022)
19. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11909–11919 (2023)
20. Wang, J., Yang, G., Chen, W., Yi, H., Wu, X., Lin, Z., Lao, Q.: Mlae: Masked lora experts for visual parameter-efficient fine-tuning. arXiv preprint arXiv:2405.18897 (2024)
21. Wang, X., Chen, T., Ge, Q., Xia, H., Bao, R., Zheng, R., Zhang, Q., Gui, T., Huang, X.J.: Orthogonal subspace learning for language model continual learning. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 10658–10671 (2023)

22. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision. pp. 631–648. Springer (2022)
23. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)
24. Wu, Y., Piao, H., Huang, L.K., Wang, R., Li, W., Pfister, H., Meng, D., Ma, K., Wei, Y.: Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. International Conference on Learning Representations (2025)
25. Zhang, J., Gu, R., Xue, P., Liu, M., Zheng, H., Zheng, Y., Ma, L., Wang, G., Gu, L.: S3r: Shape and semantics-based selective regularization for explainable continual segmentation across multiple sites. IEEE Transactions on Medical Imaging **42**(9), 2539–2551 (2023)
26. Zhang, J., Pei, J., Xu, D., Jin, Y., Heng, P.A.: Dc^2t: Disentanglement-guided consolidation and consistency training for semi-supervised cross-site continual segmentation. IEEE Transactions on Medical Imaging **44**(2), 903–914 (2024)