Bootstrap Your Own Classifier: Your Pretrained Vision Models are Secretly Strong Continual Learners

Supplementary Material

This supplementary material is organized into three main sections. First, we offer a detailed analysis and empirical validation of our proposed semantic initialization strategy. Second, we provide comprehensive details regarding our experimental setup, including dataset specifics and implementation choices, to ensure full reproducibility. Finally, we present an extensive suite of additional ablation studies that dissect the contributions of each component within our framework and justify our key design decisions.

A. Detailed Analysis of Semantic Initialization

We provide additional technical details supporting our semantic initialization strategy presented in the introduction.

A.1. Experimental Setup for Initialization Analysis

To systematically investigate the impact of classifier initialization on backbone stability and generalization, we conducted controlled experiments comparing random initialization against our proposed prototype initialization. We jointly fine-tuned pretrained backbones with their respective classifiers on three datasets: CIFAR-100, ImageNet-R, and ImageNet-A. All experiments used the same hyperparameters, with the only difference being the classifier initialization strategy.

For prototype initialization, we computed class prototypes by extracting features from the training set using the pretrained backbone f_{θ_0} and calculating the mean representation per class:

$$p_c = \frac{1}{|\mathcal{D}_c|} \sum_{(x,y)\in\mathcal{D}_c} f_{\theta_0}(x), \tag{1}$$

where $\mathcal{D}_c = \{(x, y) \in \mathcal{D} \mid y = c\}$ denotes examples of class c.

A.2. Backbone Stability Analysis

We measured backbone parameter drift by computing the L1 distance between the pretrained parameters θ_0 and fine-tuned parameters θ_{ft}:

$$\Delta = ||\theta_{\text{ft}} - \theta_0||_1. \tag{2}$$

As shown in Fig. 2a, random initialization consistently induced substantially larger parameter changes compared to prototype initialization. The magnitude of this difference—approximately 2.5× on CIFAR-100 and ImageNet-R, and 3.5× on ImageNet-A—suggests that much of the optimization effort under random initialization is spent compensating for poor initial classifier positions rather than learning task-specific adaptations.

A.3. Zero-shot Probe for Representation Quality

To assess whether the reduced parameter drift translates to better preservation of the backbone's general capabilities, we evaluated zero-shot performance on two held-out datasets: CUB-200 and Stanford Cars. These datasets were never seen during fine-tuning and thus serve as probes for the backbone's representation quality.

For zero-shot evaluation, we extracted class prototypes from the target dataset using the fine-tuned backbone and used these prototypes as a non-parametric classifier. This evaluation strategy directly measures the backbone's ability to produce discriminative features for novel classes.

The results reveal a consistent pattern:

Random initialization: Zero-shot performance degraded across all fine-tuning scenarios, indicating that the backbone adaptations were overly specialized to the fine-tuning task.

Prototype initialization: Zero-shot performance was maintained or even improved, particularly after fine-tuning on ImageNet-R and ImageNet-A, suggesting that the backbone learned broadly beneficial refinements.

A.4. Conceptual Rationale

The dramatic difference in backbone stability can be understood through the lens of optimization dynamics. With random initialization, the initial loss is dominated by the classification error from arbitrarily positioned classifier weights. The resulting gradients are large and primarily serve to move classifiers toward sensible positions. These large gradients propagate through the backbone, causing substantial parameter updates that may not be beneficial for representation quality.

In contrast, prototype initialization places classifiers near their optimal positions from the start. The initial loss is lower and more balanced between improving task-specific performance and refining representations. This leads to smaller, more targeted gradient updates that preserve the backbone's pretrained structure while allowing necessary adaptations.

B. More Details on Experimental Setup

To ensure the reproducibility of our results, this section provides a comprehensive overview of the experimental environment. We detail the five benchmark datasets used for evaluation and outline the specific implementation choices, including hardware, software libraries, and crucial hyperparameters that underpin our experiments.

Table A1. Component ablation on ImageNet-A, CUB-200, and Stanford Cars. Semantic initialization consistently provides substantial improvements over baseline fine-tuning across diverse dataset types, with the complete BYOC framework achieving optimal performance.

	Method	ImageNet-A		CUB-200		Standford Cars	
		last	*avg*	*last*	*avg*	*last*	*avg*
Baseline	Finetune	26.49	41.81	45.79	68.30	28.04	53.76
Ours	w/ Semantic initialization	57.86	66.78	86.83	92.36	62.10	73.83
	w/ Semantic initialization + EMA	60.92	69.24	86.98	92.54	65.66	76.49
	w/ Semantic initialization + calibration	59.28	68.77	87.03	92.46	64.89	75.63
	w/ all (a.k.a BYOC)	**62.82**	**71.31**	**87.14**	**92.85**	**68.82**	**77.89**

B.1. Datasets

We provide comprehensive details on the five benchmark datasets used in our experiments:

CIFAR-100 [6] contains 60,000 32×32 color images across 100 classes, with 500 training and 100 test images per class. The low resolution and limited training data per class present unique challenges for continual learning, particularly in maintaining discriminative features across sequential tasks.

ImageNet-R (ImageNet-Rendition) [2] comprises 30,000 images from 200 ImageNet classes rendered in various artistic styles including art, cartoons, deviantart, embroidery, graphics, origami, paintings, patterns, plastic objects, plush objects, sculptures, sketches, tattoos, and video game renditions. This dataset tests model robustness to domain shift while maintaining semantic content.

ImageNet-A (ImageNet-Adversarial) [3] contains 7,500 natural adversarial examples across 200 ImageNet classes. These images are naturally occurring examples that consistently fool ImageNet-trained classifiers, providing a challenging test of model robustness during continual adaptation.

CUB-200 (Caltech-UCSD Birds-200) [9] includes 11,788 images of 200 North American bird species. Following SLCA [11], we use the original train/test split rather than SimpleCIL [13] configuration, resulting in approximately 30 training images per class (50:50 train/test ratio). This limited training data tests the efficiency of continual learning methods in low-data regimes.

Stanford Cars [9] contains 16,185 images of 196 car models, spanning years 1990-2012. Classes are typically defined by make, model, and year (e.g., "2012 BMW M3 coupe"). With 10-task splitting, the final task contains only 16 classes due to dataset size constraints.

B.2. Implementation

All experiments are conducted on single NVIDIA RTX 3090 GPUs (24GB VRAM) to ensure consistency. We utilize `bfloat16` precision for computational efficiency without sacrificing accuracy. Our implementation uses PyTorch 2.6.0 [7], with pretrained models loaded via `timm` 1.0.15 [10]. Specifically, we employ ViT-B/16 pretrained on ImageNet-21K with augmentation regularization [1] (model ID: `vit_base_patch16_224.augreg_in21k`).

For reproducibility, we fix the random seed to 42 across all operations including class order shuffling, dataloader sampling, and parameter initialization. Following prototypical classification conventions [8, 12, 13], we implement cosine classifiers but with a learnable temperature parameter (initialized at 0.1) applied before softmax to improve convergence stability.

Learning rate selection proved critical: we set 0.005 for the full BYOC method but halve it (0.0025) when training without EMA. This adjustment compensates for EMA's stabilization effect, which naturally slows convergence. We employ SGD with momentum 0.9 and weight decay 0.0001 across all methods. Learning rate scheduling proved detrimental for several baseline methods, so we maintain constant learning rates throughout training for fair comparison.

Data normalization follows the pretrained model's protocol with mean and standard deviation of (0.5, 0.5, 0.5) for all channels [1]. While SimpleCIL [13] omits normalization and achieves improved zero-shot performance on natural image datasets (CIFAR-100, CUB-200), this comes at the cost of degraded performance on distribution-shifted datasets (ImageNet-R, ImageNet-A). We prioritize generalization by maintaining the original pretraining normalization scheme.

C. Additional Ablation Studies

To comprehensively validate our design choices and understand the interplay between BYOC and existing continual learning paradigms, we conduct extensive ablation studies. We first examine component contributions across diverse datasets beyond our main benchmarks, then investigate the importance of parameter learnability in our framework. Finally, we analyze BYOC's compatibility with parameter-efficient fine-tuning methods to understand whether semantic initialization benefits extend to constrained optimization settings.

Extended component analysis (Tab. A1). We extend our ablation study to ImageNet-A, CUB-200, and Stanford Cars to validate the generalizability of our findings. Semantic

Table A2. Impact of component learnability on continual learning performance. Checkmarks indicate trainable components, while crossmarks indicate non-trainable components. Results demonstrate that backbone adaptation is crucial for performance, while joint training of backbone and classifier with semantic initialization yields the best results. Classifier calibration requires learnable classifiers as fixed prototypes cannot be calibrated.

Method	**Backbone**	**Classifier**	**CIFAR-100**		**ImageNet-R**	
			last	*avg*	*last*	*avg*
w/ Semantic initialization	✗	✗	80.07	84.87	52.78	57.89
	✗	✓	80.14	84.90	52.81	57.95
	✓	✗	88.83	92.93	77.54	81.96
	✓	✓	89.92	93.23	79.05	83.79
w/ Semantic initialization + EMA	✓	✗	89.78	93.36	78.56	82.94
	✓	✓	90.93	94.04	79.69	84.53
w/ Semantic initialization + calibration	✓	✓	90.42	93.69	79.33	84.19
w/ all (a.k.a BYOC)	✓	✓	**91.45**	**94.34**	**80.47**	**85.29**

initialization provides dramatic improvements across all datasets, with gains ranging from 31.37% (ImageNet-A) to 41.06% (CUB-200) in last-task accuracy over standard fine-tuning. The addition of EMA consistently improves performance, with particularly notable gains on Stanford Cars (+3.56% last accuracy). The complete BYOC framework achieves optimal results across all datasets, demonstrating that our alignment mechanisms provide universal benefits regardless of dataset characteristics or difficulty.

Component learnability analysis (Tab. A2). To dissect the contributions of learnable parameters, we systematically freeze different components. When both backbone and classifier are frozen (first row), performance matches SimpleCIL exactly, as expected. Introducing a learnable classifier alone provides negligible improvement (+0.07% on CIFAR-100), confirming that classifier adaptation without backbone updates offers limited benefit.

The critical role of backbone adaptation becomes evident when comparing frozen versus learnable backbones (with frozen classifier): enabling backbone updates improves last-task accuracy by 8.76% on CIFAR-100 and 24.76% on ImageNet-R. Joint training of both components yields additional gains, validating our design choice. Notably, classifier calibration requires learnable classifiers to establish meaningful weight transformations—when classifiers remain as fixed prototypes, no calibration is possible. This interdependence underscores the importance of our holistic approach to continual learning.

Compatibility with PEFT (Tab. A3). We investigate whether semantic initialization benefits extend to parameter-efficient finetuning (PEFT) methods by testing visual prompt tuning (VPT) [5] and low-rank adaptation (LoRA) [4]. For standard fine-tuning without semantic initialization, PEFT methods indeed improve stability—VPT achieves 85.11% versus 76.84% baseline on CIFAR-100. However, this advantage diminishes or reverses with semantic initialization: full fine-tuning with semantic initialization (89.92%) substantially outperforms VPT with semantic initialization (83.23%).

Table A3. Compatibility of BYOC with parameter-efficient fine-tuning (PEFT) methods. Semantic initialization (sem. init.) eliminates the need for parameter restrictions—full fine-tuning with BYOC outperforms PEFT methods. VPT's random initialization conflicts with semantic alignment, while LoRA shows modest compatibility but remains suboptimal.

Method	**CIFAR-100**		**ImageNet-R**	
	last	*avg*	*last*	*avg*
Finetune	76.84	87.29	61.30	73.82
w/ sem. init.	89.92	93.23	79.05	83.79
w/ BYOC (reported)	**91.45**	**94.34**	**80.47**	**85.29**
VPT [5]	85.11	89.47	65.34	69.87
w/ sem. init.	83.23	87.90	57.94	64.35
w/ BYOC	83.44	87.07	58.12	64.47
LoRA [4]	85.71	91.62	75.12	79.95
w/ sem. init.	88.67	92.66	76.14	81.64
w/ BYOC	90.23	93.47	77.25	82.27

This reversal reveals a fundamental incompatibility between VPT and our approach. VPT's performance actually degrades with semantic initialization (-1.88% on CIFAR-100, -7.40% on ImageNet-R), likely due to initialization conflicts. VPT's randomly initialized prompts disrupt the carefully established semantic alignment between backbone and classifier at initialization, negating the stabilization benefits. In contrast, LoRA shows better compatibility due to its residual design—initialized with zero down-projection matrices, LoRA begins with no modification to

the original weights, preserving semantic alignment. Consequently, LoRA benefits modestly from semantic initialization (+2.96% on CIFAR-100) and BYOC further improves performance, though full fine-tuning remains optimal. These results suggest that BYOC's semantic initialization provides sufficient stability to unlock the full potential of parameter updates, making restrictive PEFT methods unnecessary and potentially counterproductive.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[2] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021. 2

[3] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[5] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 3

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images., 2009. 2

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 2

[8] Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. MOS: Model surgery for pre-trained model-based class-incremental learning. In *AAAI*, pages 20699–20707, 2025. 2

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. CUB-200: A dataset for fine-grained bird classification. Technical report, 2011. 2

[10] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 2

[11] Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. SLCA: Slow learner with classifier alignment for continual learning on a pre-trained model. In *ICCV*, pages 19148–19158, 2023. 2

[12] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *CVPR*, pages 23554–23564, 2024. 2

[13] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *IJCV*, 133(3):1012–1032, 2025. 2