DIET-CP: Lightweight and Data Efficient Self Supervised Continued Pretraining

Bryan Rodas[1*], Natalie Montesino[2*], Jakob Ambsdorf[3*], David Klindt[4], Randall Balestriero[5]

[1]Fordham University, [2]Rutgers Univeristy, [3]University of Copenhagen, [4]Cold Spring Harbour Laboratory, [5]Brown University

DIET for continued pretraining

Continued pretraining (CP) is a promising method to adapt foundation models to new domains. However, **current methods face two issues:**

I. In niche applications, **available datasets are often small**, limiting the applicability of SSL methods developed for large-scale pretraining, and making hyperparameter search infeasible.
II. **Models are usually released as backbone-weights only**, lacking important information to continue pretraining.

We propose to bridge this gap with **DIET-CP**, a simple CP strategy, where any strong foundation model can be steered towards a data distribution of interest.

Method

DIET: A super simple SSL method that excels on small datasets

- Requires **only 1000 images and no labels**
- **Very simple objective:** No more hyperparameters than supervised finetuning
- **Stable** across data modalities and backbones
- Provides a **significant performance boost for SOTA models** such as DINOv3

We propose adapting the representations of a foundation model in an SSL setting using cross entropy on the Datum IndEx as Target for Continual Pretraining (DIET-CP) [1]. The formulation for a backbone f_θ is as follows:

$$\mathcal{L}_{\text{DIET}}(\boldsymbol{x}_n) \;=\; \text{XEnt}\big(\boldsymbol{W}\, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n),\, n\big),\; \boldsymbol{x}_n \in \mathbb{R}^D,$$

where n is the one-hot encoded *index* of each data point in a dataset of size D. W represents a linear classification head for the DIET loss. This simple objective is an effective pretraining strategy for small datasets.

Datasets

Out-of-domain and fine-grained in-domain

MedMNIST: A number of real-life medical classification datasets across different domains. These are out-of-domain for backbone foundation models [2].

Galaxy10-DECals: Out-of-domain optical telescope data for galaxy morphology classification [3].

FGVC-Aircraft, Food101: Fine-grained in-domain classification on natural images [4,5].

Adapt any strong foundation model to *your* data distribution

… using only 1000 images and no labels!

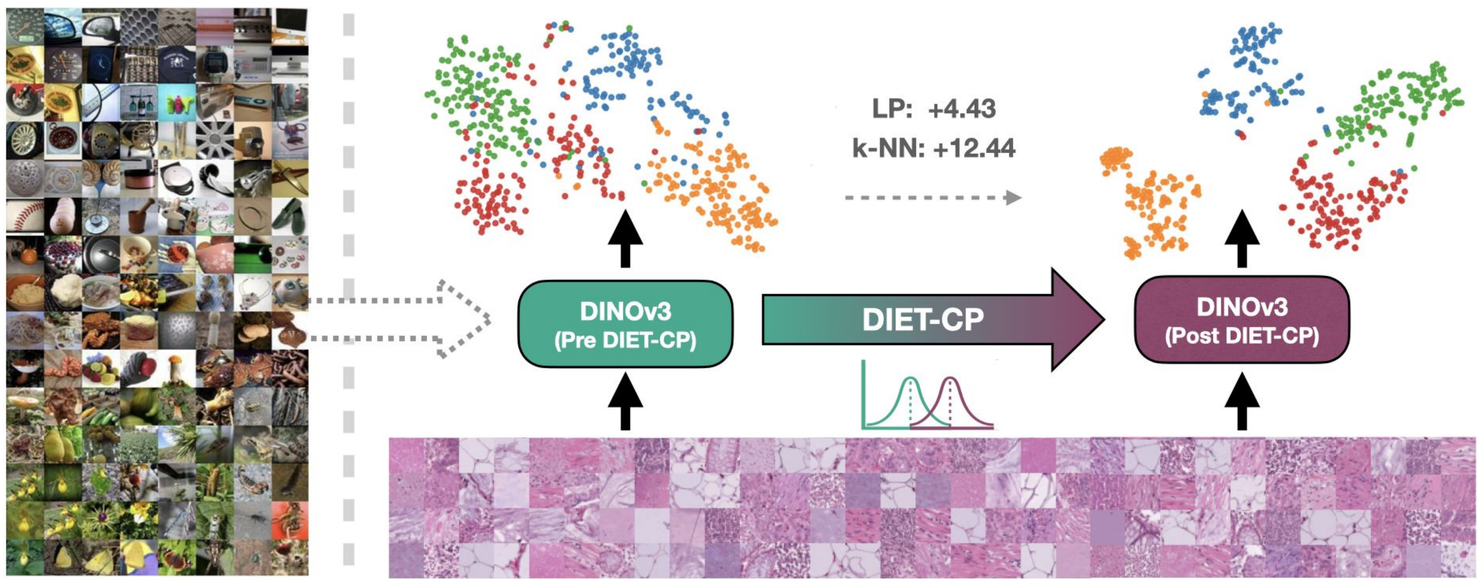

Fig. 1: DIET-CP is a label-free and efficient method for steering foundation models towards a data distribution of interest, improving class separability in the embedding space and leading to improved unsupervised and linear probing performance. t-SNE plots are generated from a PathMNIST subset.

DIET-CP Improves performance on out-of-domain datasets

On medical and galaxy morphology classification

- **DINOv2, DINOv3** [6,7] **improve by a clear margin** on k-NN and linear probing performance
- **MAE** [8]**:** weaker baseline, but larger improvements

Table 1: On average, DIET-CP improves DINOv3's linear probing (LP) by 4.43 abs. percent on F1, and on k-NN by 12.44 on MedMNIST. MAE is a weaker baseline, but benefits strongly from DIET-CP.

Backbone	Dataset	Pre DIET-CP (F1) k-NN	Pre DIET-CP (F1) LP	Post DIET-CP (F1) k-NN	Post DIET-CP (F1) LP
DINOv2	BreastMNIST	64.89	82.21	88.54 (+23.66)	88.90 (+6.69)
	DermaMNIST	21.13	40.45	41.85 (+20.72)	53.21 (+12.76)
	OCTMNIST	41.57	71.05	74.89 (+33.32)	85.41 (+14.37)
	OrganaMNIST	57.17	78.51	72.37 (+15.20)	80.30 (+1.79)
	OrgancMNIST	58.30	76.49	72.40 (+14.10)	79.02 (+2.53)
	OrgansMNIST	46.74	62.47	57.46 (+10.72)	62.21 (-0.26)
	PathMNIST	84.15	93.17	94.53 (+10.38)	95.94 (+2.77)
	PneumoniaMNIST	63.67	89.29	93.43 (+29.75)	95.93 (+6.64)
	RetinaMNIST	39.91	50.05	41.95 (+2.04)	46.06 (-3.99)
	Average	53.06	71.52	70.82 (+17.77)	76.33 (+4.81)
DINOv3	BreastMNIST	72.40	81.92	87.80 (+15.40)	91.78 (+9.86)
	DermaMNIST	22.50	47.26	33.92 (+11.42)	50.52 (+3.26)
	OCTMNIST	47.77	75.44	73.58 (+25.82)	85.02 (+9.58)
	OrganaMNIST	71.53	87.00	80.74 (+9.20)	88.33 (+1.33)
	OrgancMNIST	70.48	78.06	77.61 (+7.14)	84.57 (+6.50)
	OrgansMNIST	60.21	64.15	67.44 (+7.23)	71.95 (+7.81)
	PathMNIST	86.34	93.88	93.35 (+7.01)	95.30 (+1.41)
	PneumoniaMNIST	73.38	91.72	92.68 (+19.31)	96.08 (+4.36)
	RetinaMNIST	38.85	53.52	48.27 (+9.41)	49.25 (-4.27)
	Average	60.38	74.77	72.82 (+12.44)	79.20 (+4.43)
MAE	BreastMNIST	59.33	77.11	75.76 (+16.43)	78.46 (+1.35)
	DermaMNIST	22.90	33.23	30.43 (+7.52)	39.87 (+6.64)
	OCTMNIST	31.79	46.49	48.81 (+17.02)	66.92 (+20.44)
	OrganaMNIST	52.98	69.37	72.31 (+19.33)	78.69 (+9.32)
	OrgancMNIST	45.58	64.88	64.05 (+18.47)	71.17 (+6.28)
	OrgansMNIST	38.37	48.94	51.95 (+13.58)	60.98 (+12.04)
	PathMNIST	73.01	85.24	87.51 (+14.50)	91.76 (+6.52)
	PneumoniaMNIST	83.93	88.92	92.85 (+8.92)	93.34 (+4.42)
	RetinaMNIST	25.06	31.22	34.66 (+9.61)	39.63 (+8.41)
	Average	48.10	60.60	62.04 (+13.93)	68.98 (+8.38)

- **Fine-grained *in-domain*** remains challenging: DINO models are already strong compared to MAE and deteriorate after CP.
- DIET-CP **consistently improves models on Galaxy10-DECaLS**, a non-medical out-of-domain dataset

Table 2: Linear probing and k-NN classification performance (F1) before and after DIET-CP for non-medical datasets. FGVC-Aircraft and Food101 are considered *in-domain* fine-grained visual categorization tasks, while Galaxy 10-DECaLS is an *out-of-domain* optical telescope imaging task.

Backbone	Eval (F1)	FGVC-Aircraft Pre	FGVC-Aircraft Post	Food-101 Pre	Food-101 Post	Galaxy10-DECaLS Pre	Galaxy10-DECaLS Post
DINOv2	k-NN	19.59	30.91 (+11.31)	58.64	60.33 (+1.69)	30.53	58.30 (+27.77)
	LP	43.47	38.47 (-5.00)	73.54	65.29 (-8.25)	49.30	64.31 (+15.01)
DINOv3	k-NN	38.91	31.83 (-7.08)	63.37	58.03 (-5.34)	42.45	52.09 (+9.64)
	LP	61.00	48.56 (-12.44)	77.58	68.98 (-8.60)	57.43	62.98 (+5.54)
MAE	k-NN	3.74	6.83 (+3.09)	3.73	11.92 (+8.19)	20.44	33.93 (+13.49)
	LP	6.77	11.54 (+4.77)	10.41	21.10 (+10.69)	26.98	38.94 (+11.96)

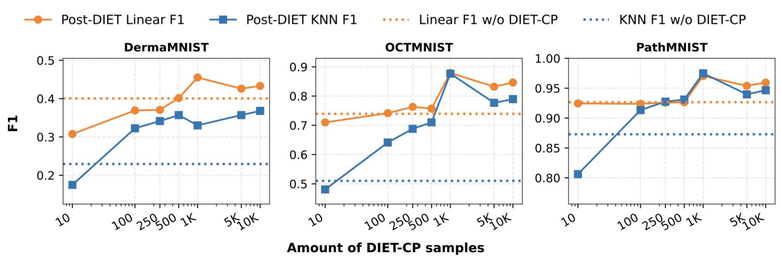

Fig. 3: Ablation study on the number of samples used for DIET-CP on a DINOv2 ViT-S. For training the k-NN and LP classifiers, the training set is kept constant with 1000 labels.

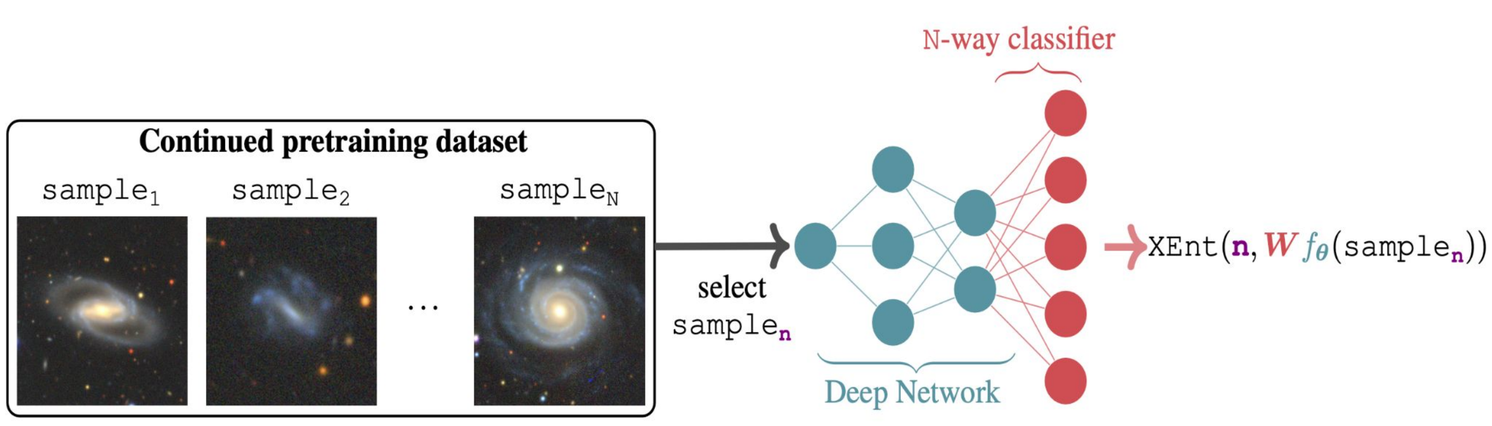

Fig. 2: DIET[1] uses the datum index (n) as the class-target, effectively turning unsupervised learning into a supervised learning problem. A standard cross-entropy loss (X-Ent) is used, with no extra care needed to handle different datasets or architectures. As opposed to other SSL methods, we do not rely on a projector nor positive views, i.e. no changes are needed compared to any existing supervised pipeline.

References

[1] Mark Ibrahim, David Klindt, and Randall Balestriero (2014). “Occam’s razor for self supervised learning: What is sufficient to learn good representations?” arXiv preprint arXiv:2406.10743.
[2] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, et al. (2013). "MedMNIST v2-A large-scale lightweight benchmark for 2D and 3D biomedical image classification." Scientific Data
[3] Mike Walmsley, Chris Lintott, Tobias Géron, Sandor Kruk, Coleman Krawczyk, et al. (2022). “Galaxy zoo decals: Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies.” Monthly Notices of the Royal Astronomical Society
[4] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi (2013). “Fine-grained visual classification of aircraft.” arXiv preprint arXiv:1306.5151
[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool (2014). “Food-101: Mining discriminative components with random forests.” ECCV
[6] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, et al. (2023). “Dinov2: Learning robust visual features without supervision.” arXiv preprint arXiv:2304.07193, 2023.
[7] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, et al. (2025) “Dinov3.” arXiv preprint arXiv:2508.10104,
[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (2022). “Masked autoencoders are scalable vision learners.” CVPR

NEURAL INFORMATION PROCESSING SYSTEMS

*Equal contribution