

Frozen Visual Representations with Pooled Analytic Readouts: A Continual-Classification Case Study

Daniel Schmitter

Abstract

A fixed feature map permits exact accumulation of a squared-error classification objective without retaining individual examples. We study this mechanism using archived Split-CIFAR-100 results and source-level reconciliation of their information and resource assumptions. On cached DINOv2 ViT-B/14 features, a cosine random-feature readout reaches mean final accuracy 0.8910 across three recorded seeds, compared with 0.8847 for linear ridge on the same frozen features. The accumulated and joint ridge implementations have identical saved accuracies, while a sequential cross-entropy head reaches 0.6470. These are not matched objectives or representations, and the experiment standardizes features using the complete training set before task arrival. It therefore establishes an exploratory final-estimation comparison under globally prepared fixed features, not a strictly online acquisition protocol or a universal no-forgetting guarantee. We derive the valid pooled-estimator identity, quantify its storage cost, and provide a small independent algebra check. The main empirical contribution is the separation of inherited representation quality, feature expansion, and evidence accumulation.

1 Introduction

Continual classification combines two difficult problems: acquiring useful visual features and updating decisions as supervision arrives. A frozen pretrained encoder separates them. If its output already distinguishes the relevant objects, a comparatively simple readout can assimilate additional labels. The resulting system may be valuable without learning a new visual representation at every task.

This paper examines a retained experiment through that separation. Its contribution is a complete account of two named empirical panels, a precise sufficient-statistic model, and qualification of what their apparent retention means. We do not claim to invent ridge regression, random features, or analytic continual classification. Nor do we interpret a strong pretrained backbone as computationally free.

2 Related work

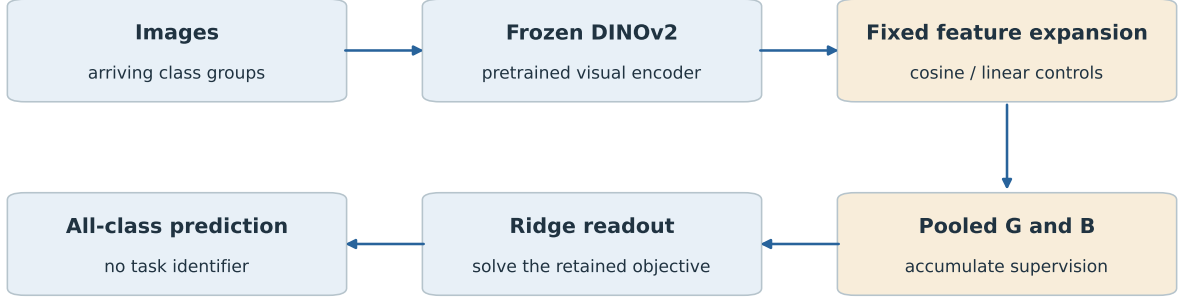
RanPAC combines pretrained representations, frozen random projections, and accumulated class statistics for continual learning [1]. Our fixed-feature readout belongs to that established family. Cosine random features have a classical kernel-approximation interpretation [2]. DINOv2 supplies the pretrained visual representation used here [3]. No backbone training or competitive reproduction of the complete RanPAC system is performed in this study.

3 Architecture and information flow

The architecture has a frozen perception stage and a fitted decision stage. Images must still pass through the pretrained encoder at inference. Standardization and the random projection are fixed before objective accumulation; changing either would invalidate old statistics. Each arriving class contributes feature cross-products and class right-hand sides in one common label schema. Solving the resulting system updates a multiclass readout without replaying individual examples.

The equality with joint ridge concerns an estimator, not a trajectory of unchanged decisions. A later class can move a decision boundary even though all earlier squared-loss contributions remain present. Moreover, a high-dimensional random feature bank trades a replay buffer for a dense quadratic object. Its memory scales quadratically in feature count and need not be small. The raw-feature ridge ablation separates the representation already supplied by pretraining from the extra nonlinear expansion.

Frozen perception, changing evidence



Read left to right above, then right to left below. Arrows show data or computation flow.

Figure 1: Frozen perception, changing evidence. Boxes distinguish supplied information, fitted components, and the quantity evaluated. Arrows show computation or data dependence, not a newly trained deep network.

4 Fixed-feature estimator

Let each task provide a feature matrix Z_t and one-hot targets Y_t in a common class schema. The feature map, centering, scaling, random projection, and regularization coefficient are held fixed. The accumulated statistics and their solution are

$$G = \sum_t Z_t^T Z_t, \quad B = \sum_t Z_t^T Y_t, \quad (1)$$

$$W^* = \operatorname{argmin}_W \sum_t \|Z_t W - Y_t\|_F^2 + \lambda \|W\|_F^2, \quad (2)$$

$$(G + \lambda I)W^* = B. \quad (3)$$

For positive lambda the normal matrix is positive definite. Expanding the finite sum of squared errors yields the normal equations, proving equality with the joint-data ridge estimator in exact arithmetic. Task order does not enter those sums. Numerical summation and solving can still depend on order. The implementation uses a linear solve rather than forming an explicit inverse.

This identity preserves an objective, not each earlier prediction. With scalar feature one, ridge coefficient one, and first target one, the fitted weight is one half. Adding target minus one changes the exact pooled weight to zero, increasing the first example’s squared error from one quarter to one. Nothing was lost from the statistics. New contradictory evidence changed the optimum. The identity also fails as a representation of old data if the feature map or its normalization changes without the necessary cross-statistics.

With d feature coordinates and C classes, dense binary32 storage for G and B is $4(d^2 + dC)$ bytes, excluding a factorization, readout, encoder, and temporary features. At the runner’s default 10,000 random features plus a constant and 100 classes, G alone occupies 400,080,004 bytes and B another 4,000,400 bytes. Raw-example-free is not synonymous with low-memory or private; the sufficient statistics may reveal individual information in small-data regimes.

5 Experimental methods

We analyze the complete three-seed ViT-B feature panel and the corresponding four-way readout ablation. The named dataset is CIFAR-100 with ten groups of ten classes, and final predictions select among all 100 classes without a task identifier. Seed-dependent class permutations and random features are generated together. No classification is rerun, no feature cache is copied, and no independent test split is created from previously exposed data.

The available runner loads cached features and computes their mean and standard deviation over the entire training set before partitioning tasks. Thus future task inputs influence preprocessing. This is global training-only preparation, not use of test labels, but it is stronger information access than strictly causal online normalization. Cache fingerprints, encoder checkpoint hashes, original commands, and pretraining contamination checks are not retained in these compact results. We identify the model by the archived naming, not by a newly verified checkpoint.

The cosine expansion appends a constant to scaled cosines of Gaussian projections. The source defaults are 10,000 projections, ridge coefficient ten, and three seeds; a comment records prior parameter tuning. The comparison head is a linear classifier on the original standardized features, trained sequentially with cross-entropy and Adam, with 100 sampled updates per task by default. It therefore differs in objective, feature dimension, and evidence retention. The joint comparison instead solves the identical ridge objective with all feature rows at once.

The ablation holds the accumulation mechanism fixed and compares raw features, rectified random projections, cosine projections, and rectified projections concatenated with raw features. It does not equalize their dimensions or optimize regularization separately. A separate deterministic NumPy check uses 71 rows, nine features, three outputs, seven blocks, seed 41840, and ridge coefficient 0.7 to compare accumulated and joint solutions. That test validates algebra, not the visual benchmark.

6 Results

Table 1: All named final-accuracy results. Values are means over the three saved seed entries, not independent dataset replications.

Method	Mean accuracy	Interpretation
Cosine expansion + ridge	0.8910	Accumulated objective
Joint cosine ridge	0.8910	Same estimator, all rows at once
Nearest class mean	0.8557	Original standardized features
Sequential linear cross-entropy head	0.6470	Different loss and retention
Raw-feature ridge	0.8847	Ablation
ReLU expansion + ridge	0.8887	Ablation
ReLU + raw ridge	0.8910	Ablation

The cosine expansion adds approximately 0.63 percentage points over raw-feature ridge. The rectified-plus-raw alternative attains almost the same mean. Consequently, the large gap to sequential cross-entropy cannot be assigned primarily to a special nonlinear expansion. Strong frozen features and pooled fitting already account for most of the performance. The nearest-class mean entries are identical across seeds because the same full features are used and class relabeling does not change their geometry.

All three accumulated accuracies equal their paired joint accuracies. Four additional task permutations per seed yield zero recorded standard deviation of final accuracy. Equality of this discrete score is not evidence of bitwise-identical weights, and the runner does not save intermediate task-accuracy trajectories. The phrase joint upper bound is also too broad: the joint value is a reference for this estimator, not a theoretical ceiling on every classifier.

7 Discussion

The practical mechanism is inexpensive supervision updates conditional on an expensive, fixed representation. This is neither distillation into an independent small perception model nor new feature discovery: prediction still needs the original encoder. The dense random-feature Gram is also a substantial retained object. Claims about edge deployment must include both costs.

The experiment supports a useful empirical case for pooled readouts, but not a new continual-learning state of the art. A genuinely matched optimization study would keep the expanded features and squared loss fixed; a strictly online study would fit preprocessing only from permitted past information. Those experiments are not inferred from the existing record. The supplement preserves all named arrays, runner sources, source hashes, and the bounded identity check.

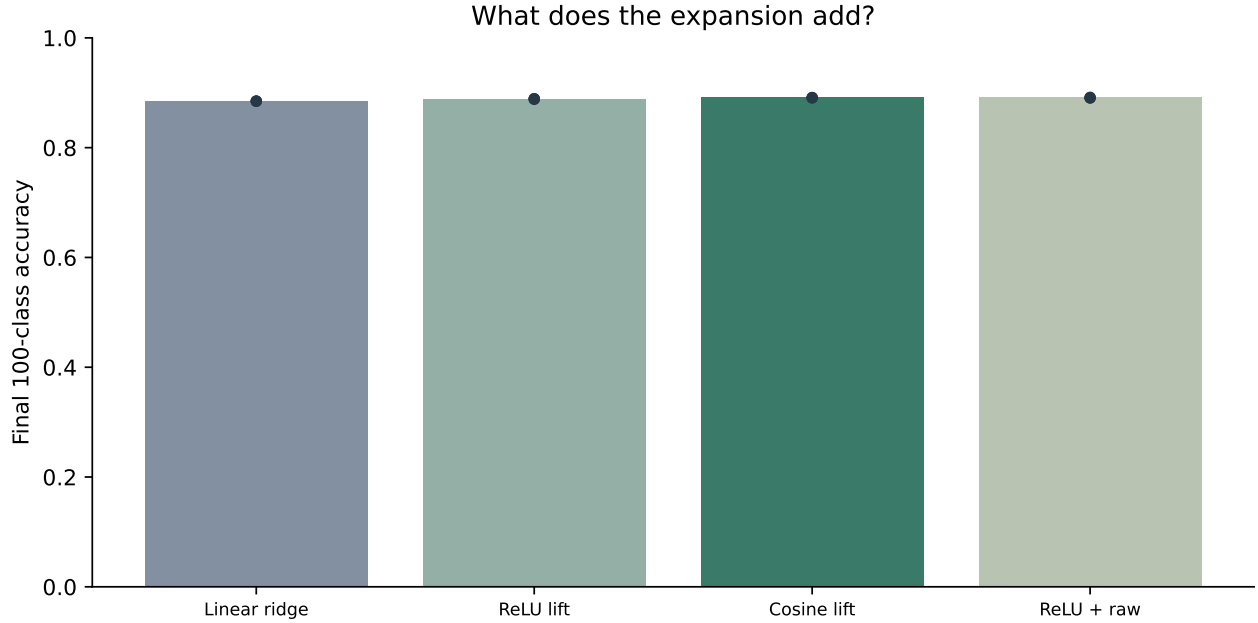


Figure 2: Complete named frozen-feature ablation, with individual seed entries. Most accuracy is already present with raw-feature ridge. The three raw-feature scores coincide; no independent data replications or confidence interval are implied.

8 Application boundary and research implication

This is a useful pattern for repeatedly updating supervision on a stable representation. It is not a demonstration that an inexpensive device acquired its own visual features or that the encoder can change without revisiting evidence. The global training-set normalization also places the retained study outside a strictly causal acquisition protocol.

9 Conclusion

Frozen representations and pooled analytic readouts can give strong final classification with no sequential readout optimizer drift. The archived result reaches 0.8910, but ordinary linear ridge already reaches 0.8847. Its scientific interpretation is an explicit decomposition of representation, expansion, and retained objective-not universal zero forgetting, a new gradient-free visual learner, or a demonstrated spline advantage.

References

- [1] M. D. McDonnell et al. RanPAC: Random Projections and Pre-trained Models for Continual Learning. NeurIPS, 2023. <https://arxiv.org/abs/2307.02251>
- [2] A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. NeurIPS, 2007. https://papers.neurips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html
- [3] M. Oquab et al. DINOv2: Learning Robust Visual Features without Supervision. Author preprint, 2023. <https://arxiv.org/abs/2304.07193>