MULTI-LABEL CLASS-INCREMENTAL LEARNING FOR OPHTHALMIC DIAGNOSIS OVER A RECURRING POPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Many real-world problems require that each example be assigned to multiple non-exclusive labels, a task known as multi-label classification. However, most of the problems considered in the context of continual learning, including the class-incremental learning (CIL) scenario, assume a single-label classification task, which does not allow an object to be assigned multiple labels simultaneously. This paper addresses that gap by introducing a multi-label incremental learning scenario that removes the assumptions of disjoint tasks and mutually exclusive classes. This scenario is particularly relevant in oculomics, which studies the relationship between ophthalmic biomarkers and systemic health or diseases. In this field, new sets of diagnoses often coexist with existing classes rather than replacing them, and re-annotation for each new problem is rarely feasible, leaving most historical images only partially labeled. We identify and formalize this scenario as a multi-label class-incremental learning problem over a recurring population, with a dynamic label space and partial relabeling, in which each label may be positive, negative, or unobserved. The label space thus expands from task to task, and training examples recur across different tasks as new sets of labels are introduced; to our knowledge, this combination has not been studied before. We evaluate the setting experimentally on the ODIR-5K dataset, a real-world multi-label ophthalmology benchmark, by instantiating the case in which every introduced label is observed, so that missingness is driven solely by the schedule. Buffer-free continual learning baselines are compared against joint and naive bounds. About half of the gap between sequential and joint training is due to forgetting, and half to lost co-training, which no buffer-free method recovers. Refitting the decision layer on an unchanged backbone recovers three-quarters of the gap — placing the residual loss in the heads rather than in the shared representation. Our findings guide the implementation and maintenance of diagnostic models whose label sets evolve.

1 INTRODUCTION

Pattern classification systems are rarely deployed in the same trained version and left unchanged. As they continue to operate, the set of categories is expected to grow clearly. A clinician begins diagnosing a condition that was previously ignored and deemed irrelevant. Retraining from scratch on the entire dataset or history is computationally wasteful, or simply impossible. Also, past data may not be retained due to data protection rules or storage limitations. The solution is to use Class-incremental learning, in which a model extends its label set on a task-by-task basis while preserving competence on previously learned classes, without storing all the data it has ever seen Zhou et al. (2024). Defining difficulty in this area is catastrophic forgetting, the tendency in a network where fine-tuning on new classes overwrites the representations used to recognize old ones. However, nearly all of this progress assumes a standard image-classification setup: tasks are disjoint, classes are mutually exclusive, and every training image is fully annotated with respect to the classes known at that time. However, none of them holds in the applications where continual learning (CL) would be most valuable.

A good example is a diagnostic model for retinal fundus images. When an ophthalmologist starts annotating a newly recognized condition, two things depart from the standard setup. First, the new

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

condition does not exclude the existing ones (diabetic and hypertensive retinopathy coexist). Thus, it adds an axis to the label space rather than to a partition region. Second, expert annotation is costly, so only a fraction of past cases receive the new label. For the rest, it is unknown, which is not the same as negative. A softmax head fits neither deviation: it cannot represent co-occurring labels, and adding a class renormalizes all existing probabilities. Filling unknown entries with negatives biases the model against the new class, and fine-tuning each task in isolation reproduces the forgetting it was meant to prevent. What fits is a set of independent binary classifiers trained with a masked loss: it grows one output at a time and keeps negatives distinct from unobserved entries. This is not merely convenient but necessary. This paper makes the following contributions:

1. We formalize multi-label class-incremental learning (ML-CIL) over a recurring population. A fixed set of images is re-annotated as the label space grows, allowing a given instance to recur across different tasks, and the labels of previously introduced classes are observed each time it does so. To our knowledge, this combination has not been studied. The existing ML-CIL literature retains the disjoint-population assumption and therefore applies its machinery to reconstructing old-class labels from a recurring population without additional effort.

2. We systematically evaluate ML-CIL on a real clinical benchmark (ODIR-5K), establishing the practical severity of catastrophic forgetting in this setting relative to standard class-incremental learning. We also quantify how much of the performance gain from joint training is attributable to forgetting versus the inherent cost of sequential rather than joint supervision.

3. We show that the order in which labels are introduced accounts for as much of the final performance as the choice of CL method, and we trace it to task size rather than position. Since joint training is invariant to the order, reversing the schedule isolates the effect of the schedule: it removes most of the forgetting component while leaving the co-training component. A further experiment that equalizes the four tasks within a fixed budget removes the effect entirely, pointing to implicit rehearsal by the last task as the mechanism.

4. We provide a comparative assessment of existing CL strategies adapted to this multi-label setting - architectural separation, buffer-free regularization, and an adapted knowledge transfer (KRT). We characterize which are feasible for clinical deployment when past patient data cannot be stored.

2 Related Work

Class-Incremental Learning. In this scenario (Van de Ven & Tolias, 2019), task identity is unavailable at test time and the classifier must cover every class seen so far (Masana et al., 2022). Remedies for catastrophic forgetting fall into three families. **Rehearsal** methods retain or revisit past data: iCaRL stores exemplars selected by herding (Rebuffi et al., 2017), and experience replay interleaves a small episodic memory with the stream (Chaudhry et al., 2019; Buzzega et al., 2020). **Regularization** methods constrain updates rather than store data: Elastic Weight Consolidation penalizes weight drift, the Fisher information marks as important (Kirkpatrick et al., 2017), and Learning without Forgetting distills the previous model's outputs on new data (Li & Hoiem, 2017). **Architectural** methods add task-specific capacity (Yan et al., 2021; Douillard et al., 2022). Almost all of this literature, including its evaluation protocols (Masana et al., 2022), assumes mutually exclusive classes and fully labeled samples; our setting drops both.

Multi-Label Classification with Incomplete Labels. Multi-label classification is implemented as a set of binary classifiers, each trained with cross-entropy as the loss function. Research on partial supervision in a fixed label space is directly relevant, and learning from only partially annotated label vectors leads to significant performance degradation Durand et al. (2019). That phenomenon was even pushed to the extreme of a single positive label per image Cole et al. (2021). The broader literature on missing labels recovers unobserved entries through label-correlation and low-ranking features Feng et al. (2020),Ding et al. (2023). We draw the masked-loss principle from the literature but combine it with a label space that grows over time - missing entries in our setup result from classes that did not yet exist when a sample was labeled, rather than from an annotation budget.

Multi-Label Class-Incremental Learning. As part of the latest, rapidly evolving research, Class-Incremental Learning is being analyzed using Multi-Label Images. *Kim et al.* first exposed the problem of imbalance in replay memories in this setting and proposed partitioning reservoir sampling Kim et al. (2020). Optimizing class distribution in memory casts memory updates as a distri-

bution matching optimization Liang & Li (2022). Recent work in this setting rests on one premise: old-class labels are absent and must be reconstructed. The mechanisms vary: an augmented label-dependency graph Du et al. (2023), dynamic pseudo-labeling with cross-attention transfer Dong et al. (2023), online relabelling with rebalanced losses Du et al. (2025), per-class thresholds taken from the previous model's confidence Zhang et al. (2025a), and example-free or prompt-based variants Zhang et al. (2025b); Ding et al. (2025); Zhao et al. (2024). Hierarchical label expansion Lee et al. (2023) is related but refines labels into fine-grained child classes rather than multi-label co-occurrence. A recurring population provides the missing labels for free, so we adapt KRT Dong et al. (2023): we drop its pseudo-label restoration while keeping its knowledge transfer. This does not make the problem easier. The old heads still receive a real gradient on a population that is no longer the one they were fitted to, which is where the residual degradation comes from (Experiment 4, Section 5).

Continual Learning in Medical Imaging. Continual learning is well motivated clinically, since privacy rules restrict data pooling and diagnostic label sets evolve Kumari et al. (2025). Medical work so far covers domain-incremental adaptation across scanners and sites Lenga et al. (2020); Srivastava et al. (2021) and single-label class-incremental streams Singh et al. (2023); Akundi & Sivaswamy (2022); Derakhshani et al. (2022). Retina is no exception: class-incremental OCT recognition Verma et al. (2023), few-shot class-incremental fundus recognition Zhang et al. (2024), and domain-incremental diabetic retinopathy detection Bravo Rocca et al. (2025); Bravo-Rocca et al. (2026) are all single-label. Oculomics makes that limiting: it treats the retina as a non-invasive window onto systemic health rather than an organ diagnosed in isolation Wagner et al. (2020), and since deep networks were shown to recover systemic risk factors from fundus photographs Poplin et al. (2018), the set of targets read from one image has kept growing Wu & Liu (2023). Such a model is not learning a sequence of disjoint tasks but expanding a shared multi-label space over a population whose earlier images stay relevant yet are only partially annotated for each new target. To our knowledge, that setting has not been explored in retinal imaging.

3 MULTI-LABEL CLASS INCREMENTAL LEARNING TASK

Let $\mathcal{X} = \{x_1, \dots, x_N\}$ be a pool of images, and annotation is a set of attributes where each attribute a with its own finite value set $\mathcal{V}_a$: binary for presence/absence ($\mathcal{V}_a = \{0, 1\}$), graded for an ordinal scale (e.g. $\{0, \dots, 4\}$ for retinopathy severity). The values are mutually exclusive within a given attribute but can co-occur across attributes, making the problem multi-label. Thus, attributes need not have the same arity. Each instance contains a complete latent annotation $y_i^\star$ assigning each attribute a a value $y_{ia}^\star \in \mathcal{V}_a$, which stands for a fixed fact about the patient.

The set of attributes is revealed incrementally as an ordered sequence of disjoint blocks $\mathcal{C}^1, \mathcal{C}^2, \dots$, with $\mathcal{C}^{1:t} = \bigcup_{s \le t} \mathcal{C}^s$. At the beginning, neither the horizon T nor the final size $|\mathcal{C}|$ is known: $\mathcal{C}^{1:t}$ is a prefix of an open-ended sequence, and the method cannot assume otherwise. Task t specifies a training subset and an observation mask $\mathcal{M}_i^t \subseteq \mathcal{C}$ of attributes visible for a given instance i: the learner sees $y_{ia}^\star$ for $a \in \mathcal{M}_i^t$ and the unobserved marker $\varnothing$ for the rest. There are mainly two reasons an entry can be missing. The mask breaks down as follows:

$$\mathcal{M}_i^t = \underbrace{\mathcal{C}^{1:t}}_{\text{schedule } \mathcal{S}^t} \cap \underbrace{\mathcal{A}_i}_{\text{availability}},$$

where $\mathcal{A}_i \subseteq \mathcal{C}$ denotes the attributes that have ever been recorded for instance i. A missing entry resulting from the fact that $a \notin \mathcal{S}^t$ is not yet introduced - identical for every instance, monotonic with respect to t, and resolved as part of a known future task. An entry that is missing because $a \notin \mathcal{A}_i$ is entered but has never been registered - instance-specific, independent of t and will never be resolved. Neither of these is a recorded negative result, and this distinction is constitutive: treating either of them as a negative result leads to false negatives occurring at the frequency of the given condition, which is most harmful in the case of rare attributes introduced last.

The instance pool is fixed, so $\mathcal{X}^t \subseteq \mathcal{X}$ and the subsets overlap: an instance positive in two blocks recurs. With full availability, the mask is shared and consistently increases as the prefix $\mathcal{M}_i^t = \mathcal{C}^{1:t}$ up to $\mathcal{M}_i^T = \mathcal{C}$. Annotation is added to known objects, as when a clinician records a new condition on archived cases.

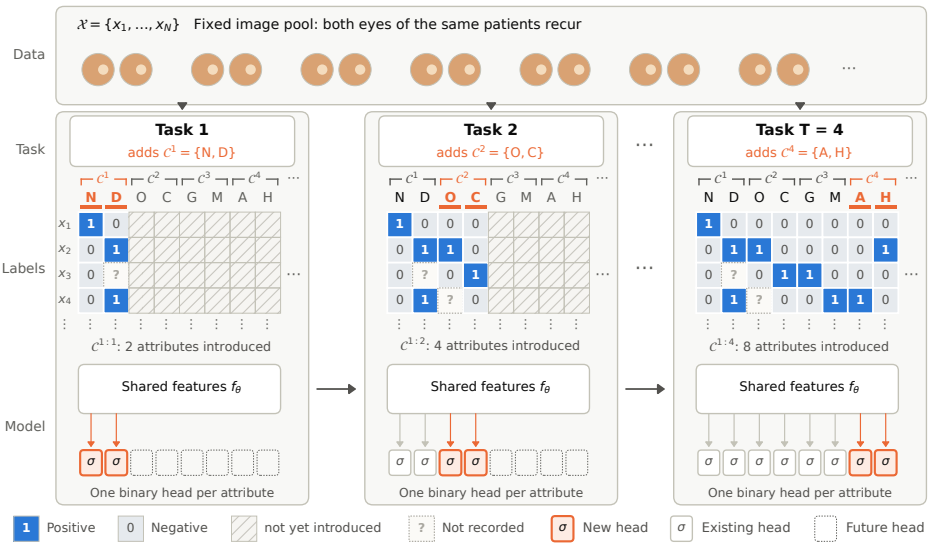

Figure 1: Multi-label class-incremental learning task concept over a recurring population.

Since attributes co-occur, the natural parametrization is not a single softmax function over $\mathcal{C}^{1:t}$, but rather an output layer for each attribute: attribute a owns a block of $|\mathcal{V}_a|$ outputs normalized within that block, ensuring that exclusivity is enforced within a given attribute. A binary attribute collapses to a single sigmoid, giving C as independent binary heads over a shared representation $\hat{y}_{ic} = \sigma(h_c(f_\theta(x_i)))$, with f_θ the backbone and h_c is the head assigned to attribute c. Task t minimizes the masked binary cross-entropy over observed entries only:

$$\mathcal{L}_t = \frac{1}{\sum_{x_i \in \mathcal{X}^t} |\mathcal{M}_i^t|} \sum_{x_i \in \mathcal{X}^t} \sum_{c \in \mathcal{M}_i^t} \mathrm{BCE}(\hat{y}_{ic}, y_{ic}^\star) \,.$$

Therefore, unobserved entries do not contribute any gradient. The evaluation takes place on a separate split that has been fully annotated in the space $\mathcal{C}$, enabling comparability of results for early and late attributes.

Classic class incremental learning is based on assumptions we do not share: 1) **Instance disjointness** - $\mathcal{X}^s \cap \mathcal{X}^t = \emptyset$ for $s \neq t$; a new task arrives on a new population 2) **Label exclusivity** - $\sum_c y_{ic}^\star = 1$; the classifier is a softmax over a partition of the label space. Here, attributes co-occur - 498 of the 7,000 images carry two or more labels. Two labels can both be positive, leaving the softmax undefined. 3) **Annotation completeness** - every training image is fully labeled with respect to the classes that are known at the time. An unobserved entry is not a recorded negative.

4 Experimental Setup

The experiments were designed to answer the following research questions:

RQ1 - How much does a dynamic label space with co-occurring diseases cause catastrophic forgetting when fine-tuning sequentially vs jointly training? Is there less interference than in classic class-incremental learning, where each task introduces new object categories instead of extra labels in partially recurring ones? Additionally, how much of that forgetting is a property of the order in which the labels arrive?

RQ2 - Can a model that keeps the shared representation fixed and learns each class through a separate decision module be competitive with models that adapt the representation at the same time keeping previous class decisions unchanged?

RQ3 - Can parameter space methods, such as Elastic Weight Consolidation, or output space methods, such as old-head anchoring, regularize the shared backbone so that it performs well on previ-

ously learned classes without storing the data visible in the past? How close do these methods come to a joint training and to a zero-forgetting architecture baseline?

RQ4 - What is the nature of the residual degradation produced by the most effective buffer-free methods? Is it in the shared representation or the decision layer, and are they overconfident and prone to false positives, as suggested by recent work on multi-label class-incremental learning in a disjoint, coupled-label setting? Together, these factors define whether the context is better described as a refined decision dilemma in a predominantly stable context or as continuous learning.

Data All experiments were conducted using the *ODIR-5K* ophthalmic benchmark[1], which contains 3,500 patients and 7,000 color fundus images (left and right eye of each patient), together with 8 labels - (N) normal, (D) diabetic, (G) glaucoma, (C) cataract, (A) AMD - Age-Related Macular Degeneration, (H) hypertensive, (M) myopia, and (O) other. Each eye is labeled from its own diagnostic keyword string rather than from the patient-level vector the dataset ships, because a patient's two eyes often disagree: in 1,271 of the 3,500 patients (36.3%) the two derived vectors differ, so a patient-level label would be wrong for at least one eye in over a third of the archive (Appendix A.1). Per-eye positives range from 3,125 (N) to 193 (H), and 498 of the 7,000 images carry two or more labels (Tab. 1).

Each of the eight annotations is a binary attribute that indicates the presence or absence of a given condition. The annotation is complete per image (6,502 images carry exactly one disease, 490 two, and 8 three), so $\mathcal{A}_i = \mathcal{C}$ and the observation mask reduces to the schedule term $\mathcal{C}^{1:t}$. The availability-driven missingness of Section 3 is not exercised here (see Limitations in Section 6).

The two labels introduced by a single task are independent attributes grouped by frequency of occurrence, rather than alternative values of a single attribute, and they do indeed occur together: 21 images (0.3%) carry both labels of the same task, which would leave a per-task softmax undefined and is the first reason each label owns an independent binary head. Appendix A.1 gives the co-occurrence counts, within a task and across tasks.

Table 1: Eight ODIR-5K classes, number of eyes positive for each, and the task that introduces it.

	N	D	O	G	C	A	M	H
condition	normal	diabetic	other	glaucoma	cataract	AMD	myopia	hypert.
images	3125	1801	1200	326	313	280	268	193
introduced at	T_1	T_1	T_2	T_3	T_2	T_4	T_3	T_4

Employed methods, architectures, measures. The *ResNet-18* architecture pre-trained on ImageNet was selected as the backbone due to its widespread use and small size. Its final layer is replaced by the C independent sigmoid heads of Section 3, one per label, and all eight ODIR classes are binary. All heads are instantiated at start-up and exposed only through the observation mask, which is equivalent to allocating them on arrival. All models are trained with the masked binary cross-entropy loss of Section 3, which supervises only the observed labels; unobserved entries contribute no gradient and are never treated as negatives.

Because past patient images may not be retained, we restrict the study to methods that do not store historical data. Within that family we order the methods along a single axis — how far the shared representation is allowed to move as new tasks arrive — which is what determines how much plasticity the network retains across tasks: joint training (upper bound, full plasticity and full supervision), *naive sequential fine-tuning* (lower bound), where the entire backbone and heads adapt each task, so interference spreads through the shared representation, *architectural separation*, in which the backbone is frozen and each task trains only the decision head of its new classes, thereby extending the structure while knowledge is reused exclusively through transfer, and past decision functions are unchanged. In SEP-IN, the backbone is a network pre-trained on ImageNet, whereas in SEP-BASE the model is fine-tuned on the first base task T_1 and then frozen. Finally, *controlled plasticity* has backbone which is trainable but is subject to a penalty via *Elastic Weight Consolidation* Kirkpatrick et al. (2017) in the parameter space and *old head anchoring* in the output space (the previous model is preserved as a frozen reference point and the current model's sigmoid outputs on old class heads are anchored to it - this is a partial constraint for each head, because the heads are independent

[1] https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

and cannot be expressed using the softmax function Li & Hoiem (2017). We additionally include *KRT* Dong et al. (2023), adapted to our scenario: we omit pseudo-label restoration for old classes while retaining its knowledge transfer. The previous model is frozen, and both head outputs and backbone features are compressed into the current one. In the calibration analysis in Experiment 4, we additionally consider the *maximum entropy* probe: the entropy regularizer from Du et al. (2024) added to our best regularizer, used solely to verify whether the calibration remedy is applicable in our case. We deliberately do not use the accompanying correlation graph mechanism, which targets the coupled label space we do not have.

Writing $\mathcal{O}^t = \mathcal{C}^{1:t-1}$ for the heads of the previously introduced classes and $(\tilde{f}, \tilde{h})$ for the frozen copy of the model saved after task $t-1$, the three penalties we add to $\mathcal{L}_t$ are, for a batch B:

$$\mathcal{L}^{\text{anc}} = \frac{\lambda_{\text{anc}}}{|B||\mathcal{O}^t|} \sum_{i \in B} \sum_{c \in \mathcal{O}^t} \text{BCE}\Big(\hat{y}_{ic},\ \sigma(\tilde{h}_c(\tilde{f}(x_i)))\Big)\,,$$

a binary cross-entropy between the current sigmoid and the frozen model's sigmoid on the old heads only, with the frozen output used as a soft target and no gradient flowing through it. KRT adds to the same term a feature-level penalty,

$$\mathcal{L}^{\text{krt}} = \mathcal{L}^{\text{anc}} + \frac{\lambda_{\text{feat}}}{|B|\,d} \sum_{i \in B} \| f_\theta(x_i) - \tilde{f}(x_i) \|_2^2,$$

the mean squared error between the current and frozen features ($d = 512$ for ResNet-18). The max-entropy probe subtracts the mean per-head binary entropy over the observed entries,

$$\mathcal{L}^{\text{ent}} = -\,\lambda_{\text{ent}}\, \frac{1}{\sum_{i \in B} |\mathcal{M}_i^t|} \sum_{i \in B} \sum_{c \in \mathcal{M}_i^t} H(\hat{y}_{ic})\,, \qquad H(p) = -p \log p - (1-p)\log(1-p),$$

so that minimizing $\mathcal{L}_t + \mathcal{L}^{\text{ent}}$ maximizes output entropy and penalizes saturated predictions. EWC uses the standard diagonal-Fisher penalty of Kirkpatrick et al. (2017). We use $\lambda_{\text{anc}} \in \{0.5, 1, 2\}$, $\lambda_{\text{feat}} = 1$, $\lambda_{\text{ent}} = 0.004$ and $\lambda \in \{10, 10^2, 10^3\}$ for EWC.

Performance is primarily measured using *mean average precision* (mAP) - the mean over classes of the per-class average precision - in accordance with the practice followed in the literature on multi-label incremental learning Du et al. (2023); Dong et al. (2023); Du et al. (2024), reported both as an average across individual tasks, and for the final one. We additionally report the F1 score for individual classes and the overall F1 (respectively, CF1, OF1), as a medical imaging complement, the area under the ROC curve (AUC), and summary calibration measures reported for every method, the false positive rate (FPR), together with expected calibration error (ECE). The results of a given run are summarized by **(i)** the final mAP averaged over the eight classes, **(ii)** *forgetting*, the mean decline of a class from its own peak, $F = \frac{1}{|\mathcal{C}^{1:T-1}|} \sum_{c \in \mathcal{C}^{1:T-1}} (\max_t a_{c,t} - a_{c,T})$, averaged over the classes introduced before the last task. Classes introduced by T itself are excluded, having no later task over which to decline; Table 2 reports both averages. **(iii)** the real-time training duration time. Since classes enter at different tasks, we further decompose the gap relative to joint training into a *forgetting* component (classes whose score declines after their task) and a *co-training* component (classes that trail joint while never declining), with the latter defining the upper bound of what any continuous method can recover.

We present precision and recall gaps broken down by class (CR$-$CP and OR-OP, following Du et al. (2024)), the average prediction entropy, and later compare these with FPR and ECE across all methods, together with the max-entropy probe. We additionally report plasticity, which shows how well new labels are acquired before any interference.

Experimental protocol The eight labels are introduced over four tasks, ordered from frequent to rare classes, T_1=$\{N, D\}$, T_2=$\{O, C\}$, T_3=$\{G, M\}$, T_4=$\{A, H\}$. Each sample is assigned to every task for which it carries at least one positive label, so a multi-label image may recur across tasks (Section 3). All splits into training, validation, and test sets ($70/15/15$) are performed for each patient, so neither eye of a single patient is ever on opposite sides of a split. The optimization uses the *AdamW* algorithm with a learning rate of 0.0001 for the backbone, and 0.001 for the head. Batch size is 32, with 50 epochs, using data augmentation consisting of horizontal flipping and mild color jitter (brightness and contrast 0.1), and images are finally resized to 224×224. Every configuration is run with three random seeds, each re-randomizing the patient split, and results are reported as the mean and standard deviation over the three runs.

5 EXPERIMENTAL EVALUATION

Experiment 1. Table 2 shows the lower and upper bounds for both naive sequential fine-tuning and joint training over three seeds, and Appendix A.2 gives every metric for every method. Under the mean average precision metric, joint training substantially outperforms naive fine-tuning, which recovers only about half of the achievable mAP (0.278 against 0.561). The size of that margin strongly depends on the metric. It is more than twice as large under mAP (0.283) as under AUC (0.127) because mAP is precision-oriented and sensitive to detecting imbalanced positive classes. It shows the degradation that AUC hides, as this is a ranking measure dominated by the negatives.

The observed gap is not attributable to catastrophic forgetting alone. Table 2 breaks down the class-specific disparity in joint training by separating it into the forgetting component and the lost co-training component (the residual disparity for classes that did not experience any decline), since sequential learning cannot jointly optimize all labels in a common representation. Expressed as contributions to the 0.283 mAP gap (summed over classes and divided by eight, as mAP itself is) the two are of comparable size, 0.151 and 0.132. For visually distinctive conditions introduced early, such as cataract and myopia, the gap is dominated by forgetting. The rare conditions behave differently: glaucoma never declines, yet reaches only 0.128 against 0.359 in the joint. AMD and hypertensive retinopathy are introduced last, so their forgetting is zero by construction; the relevant quantities are their gaps to the joint: 0.319 and 0.091. For all three, the shortfall is the absence of joint supervision, not the loss of something previously held. Since the mask is $\mathcal{M}_i^t = \mathcal{C}^{1:t}$, old and new labels do receive gradients in the same batch; what neither regularization nor architectural isolation provides is that supervision on the population the old label was originally fitted to. This is what bounds recovery during training, and Experiment 4 shows that part of the component is nevertheless recoverable afterward, by refitting the decision layer on complete labels. Rehearsal is the one family that would restore that population, and it is outside the paper's scope.

Table 2: Per-class average precision for naive fine-tuning, the head refit and joint training, with the forgetting of each class, sorted by the naive–joint gap. T is the task that introduces the class.

	Class	T	naive	refit	joint	forget
C	cataract	2	0.301±0.115	0.829±0.045	0.894±0.007	0.552±0.107
M	myopia	3	0.403±0.157	0.827±0.076	0.895±0.019	0.440±0.125
A	AMD	4	0.093±0.026	0.297±0.076	0.411±0.131	—
G	glaucoma	3	0.128±0.035	0.270±0.025	0.359±0.062	0.000±0.000
N	normal	1	0.465±0.033	0.668±0.035	0.694±0.020	0.156±0.067
O	other	2	0.234±0.038	0.305±0.009	0.450±0.024	0.000±0.000
D	diabetic	1	0.446±0.070	0.495±0.051	0.546±0.039	0.060±0.050
H	hypertensive	4	0.151±0.052	0.217±0.084	0.241±0.096	—
mean over all eight classes			0.278	0.489	0.561	—
mean forgetting over the six classes of $T_{1:3}$						0.201
forgetting as a contribution to the mAP gap						0.151

Forgetting is class-dependent rather than uniform: the normal class degrades gradually as tasks are added, and cataract declines once later conditions are learned, whereas diabetic retinopathy, which co-occurs with conditions introduced later, remains stable. Because the schedule is fixed by class frequency, we reran both bounds in reverse order with the same seeds, splits, and hyperparameters. Joint training is unaffected (0.562 against 0.561 mAP), so order is the only difference between the two runs. Naive fine-tuning rises from 0.278 to 0.397 mAP, its forgetting falls from 0.201 to 0.085, and the gap to joint narrows from 0.283 to 0.165. The change is confined to the classes whose position moved: cataract and myopia forget 0.552 and 0.440 when learned early, but only 0.182 and 0.162 when they arrive late, while normal and diabetic retinopathy now enter last and, like AMD and hypertensive retinopathy in the forward order, have no subsequent task over which to decline. The rare classes are insensitive to position — AMD, glaucoma, and hypertensive retinopathy end between 0.09 and 0.19 mAP under both schedules. Both observations follow from the same mechanism: in a recurring population, the labels of earlier classes remain observable in the revisited samples, so a large late task acts as implicit rehearsal. We adopt the frequent-to-rare order as the more demanding protocol. Appendix A.3 separates two linked factors: with equal task sizes, class order affects neither mAP nor forgetting.

Experiment 2. When comparing naive fine-tuning with frozen variants, the crucial factor to consider is not the ability to retain prior knowledge, but the degree to which a static representation can help identify newly presented classes. Both beat naive sequential fine-tuning (Appendix, Table 3): a model that freezes its representation and adds an independent head per class, storing nothing and forgetting nothing, is a stronger baseline than plastic fine-tuning. Contrary to expectation, SEP-IN on untouched ImageNet features outperforms SEP-BASE, whose backbone was first adapted to the dominant base pathologies, by $+0.039$ mAP on every seed (Appendix, Table 7). Appendix, Fig. 5 illustrates why: adapting to $T_1{=}\{\mathrm{N}, \mathrm{D}\}$ is beneficial to the base classes but specializes the representation at their expense. Myopia (M) and hypertensive retinopathy (H), introduced later, lose most of their support under SEP-BASE. A representation shaped around the dominant base classes does not transfer to the rare variant. A general-purpose representation, on the other hand, does.

Experiment 3. The two regularizer families separate cleanly (Fig. 2). Anchoring in the output space reduces the forgetting rate from 0.20 to 0.03 mAP and raises the final score from 0.278 to 0.407. KRT reaches 0.408, and the two are not separable at three seeds (a paired difference of $+0.001$, with the sign changing between seeds; Appendix, Table 7); both sit slightly above the frozen SEP-IN baseline, and that difference is consistent at $+0.010$ on every seed. EWC in parameter space does not provide any of these benefits. At its best setting ($\lambda{=}10$), it reaches 0.306 mAP with forgetting still at 0.171, and the paired difference to naive fine-tuning is $+0.029$ with the sign changing between seeds - at three seeds, EWC is not distinguishable from accomplishing nothing. The sweep in Appendix, Table 6 moves the final mAP by 0.016 across three orders of magnitude of λ, so this is insensitivity to the penalty rather than a badly chosen λ. In a setting with masked BCE, a recurring population, independent heads, and the protection weights the Fisher considers important, it does not translate into protection of predictions for old classes, whereas directly anchoring those predictions does.

The limiting factor is stability, not plasticity. All methods without a buffer (naive, EWC, anchoring, KRT) learn new classes equally well (see the plasticity in Appendix Table 3). They are distinguished only by retention, so anchoring provides stability without costing plasticity. These methods may be grouped into low-forgetting (anchoring, KRT, their combination) and high-forgetting (naive, EWC), as well as a zero-forgetting frozen architecture.

Cost. Every sequential method makes four passes over the task stream instead of one, so all of them (naive, the frozen variants, EWC, anchoring, and KRT) cost roughly twice a joint run. EWC's Fisher pass and the frozen reference pass used by anchoring and KRT add a few percent on top of it.

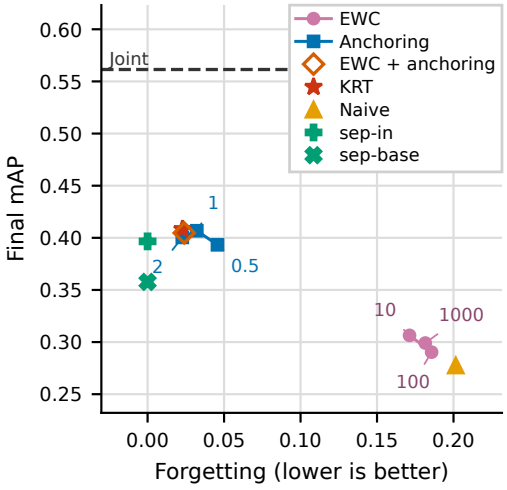

Figure 2: Final mAP against forgetting, with multiple tested methods. The EWC λ-sweep and the anchoring weight sweep are shown as connected series, labeled by their setting.

Experiment 4. The residual degradation in quality resulting from the use of buffer-free methods is analyzed at the decision-making level. A head-refit diagnostic freezes the representation of a completed naive run and retrains all eight heads on a fully labeled training set. Since the backbone never changes, the recovered result measures the misalignment in the decision layer rather than the drift of the shared representation. Refitting the heads recovers about three-quarters of the gap between the naive runs, lifting the mAP from the naive lower bound (0.278 to 0.489) against the joint value (0.561), as Figure 3a shows. That places it above every buffer-free regularizer from Experiment 3 and both frozen variants from Experiment 2, using the non-retrained backbone. The recovered performance was therefore not lost from the representation but remained in the decision layer as a result of sequential learning. The recovery is not uniform, as the refit column of Table 2 shows: the classes attributed to forgetting in Experiment 1 are almost fully restored, with cataract going from 0.30 to 0.83, myopia from 0.40 to 0.83, and normal from 0.47 to 0.67. The classes identified as lost co-training recover only in part, AMD reaches 0.30 of the joint 0.41, glaucoma 0.27 of 0.36, and other pathology 0.31 of 0.45. What remains after the refit is very small, 0.072 mAP on average. This is visible exactly on these rare

co-occurring classes - residual is the representation-related part of the gap. Appendix A.8 crosses this decomposition with the one from Experiment 1. The two are not nested: the decision-layer component (0.211) exceeds the entire forgetting component (0.151), and this excess is attributable to the four classes that never declined: refitting recovers 0.060 mAP of their gaps, which is 45% of the total co-training component. For the classes that declined, the forgetting and decision-layer components agree on 0.0005 mAP, although they are measured independently.

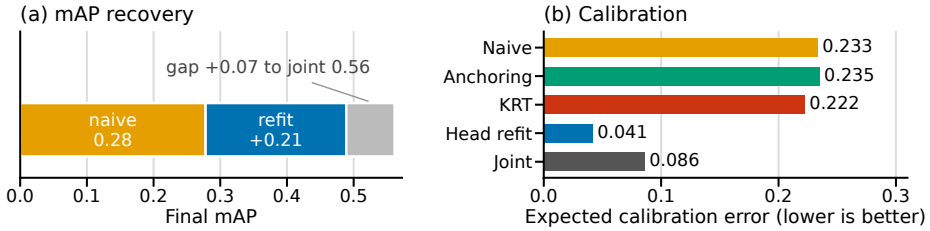

Figure 3: (a) The naive-to-joint mAP gap, split into the part the head refit recovers and the part that remains. (b) Expected calibration error of the final model for each method.

The decision layer is also miscalibrated, as recent work on multi-label class-incremental learning predicts. The naive model is over-confident: near-zero prediction entropy (0.092), a high false-positive rate (0.251), and wide recall-over-precision gaps (Appendix, Tables 3 and 4). Anchoring and KRT improve part of this picture — they lower the false-positive rate to 0.215 and 0.211, narrow both gaps, and raise prediction entropy roughly fourfold — but leave the expected calibration error where it was (0.235 and 0.222, compared with 0.233). Only the head-refit is well calibrated, at an ECE of 0.041, below even joint training (Fig. 3b): fitting heads to complete labels on a stable representation is exactly what calibration requires. Adding a maximum-entropy penalty to output anchoring, the strongest regularization baseline, produces minimal changes: the mAP value changes by less than 0.01 (0.407 to 0.406), and all calibration metrics remain within the bounds of the standard anchor's initial noise. The head rebuild takes roughly as long as a naive run plus the refit pass, and the probe with maximum entropy costs the same as an anchoring, at the naive level.

6 Conclusion

We studied multi-label class-incremental learning over a recurring patient population, in which each task adds labels to a fixed archive rather than introducing new images. Four experiments give one picture. Half of the gap to joint training is forgetting, and half is lost co-training. No buffer-free method recovers the second half during training: old and new labels do share a batch, but an old label is never supervised again on the population it was fitted to. Which half dominates depends on the schedule, not on the method. Reversing task order reduces forgetting only when the task sizes differ widely. With equal sizes, order affects neither mAP nor forgetting, so the effect follows task size rather than position. Against forgetting, only methods that act in the output space help, and they only match a frozen backbone. Freezing is not free either: a backbone fitted to the large-base task serves the rare classes worse than untouched ImageNet features do. Refitting the decision layer on an unchanged backbone recovers most of what remains and restores calibration. Therefore, the heads are mainly responsible for the loss, not the shared features. Expanding a label space over a recurring population is closer to transfer learning than to classical class-incremental learning.

Limitations. Our conclusions apply to an intentionally limited setting. We assume a closed pool of instances with full availability: each task re-annotates the same archive, and the labels of old classes are observed again whenever an instance is revisited. Therefore, a missing label means that it has not been entered yet, not that it is unknowable, so the availability term $\mathcal{A}_i$ of Section 3 is defined but never exercised. We exclude rehearsal and keep to regularization- and architecture-based methods. The head refit is a diagnostic rather than a method, since it uses the complete label matrix that no method here can see. Finally, we use one dataset (ODIR-5K), one architecture (ResNet-18) and one grouping of the labels into four tasks. We vary the order and the task sizes but not the grouping, which the size effect suggests matters as well. The formulation also allows graded attributes, which we do not test.

AI USE STATEMENT

Generative AI tools were used in the preparation of this paper for language editing of the authors' text and as a coding assistant for parts of the experimental Python scripts. All AI-assisted code was read, executed, and verified against the authors' recorded per-seed outputs. No text, result, figure, or citation in this paper was generated without author review, and the authors take full responsibility for its content. Generative AI was not used to generate research ideas, produce experimental results, or analyze those results.

ETHICS STATEMENT

This study uses ODIR-5K, a publicly released, de-identified ophthalmic dataset, under its published terms of use; no new patient data was collected, and no identifiable information was accessed. The models reported here are research artifacts and are not validated for clinical use: average precision on the rare conditions is far below what a diagnostic aid would require, and the calibration results in Experiment 4 show that a sequentially trained model of this kind is systematically over-confident on exactly the classes it has learned least well. Reporting that over-confidence, rather than only the headline mAP, is the reason the calibration diagnostics are in the main text.

REPRODUCIBILITY STATEMENT

Section 3 defines the setting and the masked objective; Section 4 gives the penalty terms with their λ values, the schedule, the patient-level $70/15/15$ split, and every optimization hyperparameter. ODIR-5K is public; Appendix A.1 describes how the per-eye labels are derived and validated against the patient-level labels the dataset ships with, and the keyword-to-class mapping is included verbatim in the supplementary code. All results are means over three seeds, each re-randomizing the patient split, with the sample standard deviation; per-seed paired differences are in Appendix Table 7. The experimental code is included in the supplementary material.

REFERENCES

Prathyusha Akundi and Jayanthi Sivaswamy. Incremental learning for a flexible cad system design. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2022.

Gusseppe Bravo-Rocca, Peini Liu, Jordi Guitart, Ajay Dholakia, David Ellison, and Rodrigo M Carrillo-Larco. Adaptive prompt evolution for continual learning in diabetic retinopathy detection. *SN Computer Science*, 7(5):457, 2026.

Gusseppe Jesus Bravo Rocca, Peini Liu, Jordi Guitart Fernández, Ajay Dholakia, David Ellison, and Rodrigo M Carrillo Larco. Experience replay and zero-shot clustering for continual learning in diabetic retinopathy detection. In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2025-Volume 2: VISAPP: Porto, Portugal, February 26-28, 2025*, pp. 81–92. SciTePress, 2025.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. *arXiv preprint arXiv:2106.09708*, 2021.

Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Tom van Sonsbeek, Xiantong Zhen, Dwarikanath Mahapatra, Marcel Worring, and Cees GM Snoek. Lifelonger: A benchmark for continual disease classification. In *International conference on medical image computing and computer-assisted intervention*, pp. 314–324. Springer, 2022.

Chenhao Ding, Songlin Dong, Zhengdong Zhou, Jizhou Han, Qiang Wang, Yuhang He, and Yihong Gong. Class-independent increment: An efficient approach for multi-label class-incremental learning. *arXiv preprint arXiv:2503.00515*, 2025.

Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. Exploring structured semantic prior for multi label recognition with incomplete labels. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3398–3407. IEEE, 2023.

Songlin Dong, Haoyu Luo, Yuhang He, Xing Wei, Jie Cheng, and Yihong Gong. Knowledge restore and transfer for multi-label class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18711–18720, 2023.

Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9275–9285. IEEE, 2022.

Kaile Du, Fan Lyu, Linyan Li, Fuyuan Hu, Wei Feng, Fenglei Xu, Xuefeng Xi, and Hanjing Cheng. Multi-label continual learning using augmented graph convolutional network. *IEEE Transactions on Multimedia*, 26:2978–2992, 2023.

Kaile Du, Yifan Zhou, Fan Lyu, Yuyang Li, Chen Lu, and Guangcan Liu. Confidence self-calibration for multi-label class-incremental learning. In *European Conference on Computer Vision*, pp. 234–252. Springer, 2024.

Kaile Du, Yifan Zhou, Fan Lyu, Yuyang Li, Junzhou Xie, Yixi Shen, Fuyuan Hu, and Guangcan Liu. Rebalancing multi-label class-incremental learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16372–16380, 2025.

Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 647–657. IEEE, 2019.

Lei Feng, Jun Huang, Senlin Shu, and Bo An. Regularized matrix factorization for multilabel learning with missing labels. *IEEE transactions on cybernetics*, 52(5):3710–3721, 2020.

Chris Dongjoo Kim, Jinseo Jeong, and Gunhee Kim. Imbalanced continual learning with partitioning reservoir sampling. In *European conference on computer vision*, pp. 411–428. Springer, 2020.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Pratibha Kumari, Joohi Chauhan, Afshin Bozorgpour, Boqiang Huang, Reza Azad, and Dorit Merhof. Continual learning in medical image analysis: A comprehensive review of recent advancements and future prospects. *Medical Image Analysis*, pp. 103730, 2025.

Byung Hyun Lee, Okchul Jung, Jonghyun Choi, and Se Young Chun. Online continual learning on hierarchical label expansion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11727–11736. IEEE, 2023.

Matthias Lenga, Heinrich Schulz, and Axel Saalbach. Continual learning for domain adaptation in chest x-ray classification. In *Medical Imaging with Deep Learning*, pp. 413–423. PMLR, 2020.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Yan-Shuo Liang and Wu-Jun Li. Optimizing class distribution in memory for multi-label online continual learning. *arXiv preprint arXiv:2209.11469*, 2022.

Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Amritpal Singh, Mustafa Burak Gurbuz, Shiva Souhith Gantha, and Prahlad Jasti. Class-incremental continual learning for general purpose healthcare models. *arXiv preprint arXiv:2311.04301*, 2023.

Shikhar Srivastava, Mohammad Yaqub, Karthik Nandakumar, Zongyuan Ge, and Dwarikanath Mahapatra. Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In *MICCAI Workshop on Domain Adaptation and Representation Transfer*, pp. 226–238. Springer, 2021.

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Tanvi Verma, Liyuan Jin, Jun Zhou, Jia Huang, Mingrui Tan, Benjamin Chen Ming Choong, Ting Fang Tan, Fei Gao, Xinxing Xu, Daniel S Ting, et al. Privacy-preserving continual learning methods for medical image classification: a comparative analysis. *Frontiers in Medicine*, 10: 1227515, 2023.

Siegfried K. Wagner, Alistair R. Denniston, and Pearse A. Keane. Insights into systemic disease through retinal imaging-based oculomics. *Translational Vision Science & Technology*, 9(2):6, Feb 2020. doi: 10.1167/tvst.9.2.6. URL https://doi.org/10.1167/tvst.9.2.6.

Jo-Hsuan Wu and Tin Yan Alvin Liu. Application of deep learning to retinal-image-based oculomics for evaluation of systemic health: A review. *Journal of Clinical Medicine*, 12(1):152, 2023. ISSN 2077-0383. doi: 10.3390/jcm12010152. URL https://www.mdpi.com/2077-0383/12/1/152.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3013–3022. IEEE, 2021.

Aoting Zhang, Dongbao Yang, Chang Liu, Xiaopeng Hong, and Yu Zhou. Specifying what you know or not for multi-label class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22345–22353, 2025a.

Jinghua Zhang, Peng Zhao, Yongkun Zhao, Chen Li, and Dewen Hu. Few-shot class-incremental learning for retinal disease recognition. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Xiang Zhang, Run He, Jiao Chen, Di Fang, Ming Li, Ziqian Zeng, Cen Chen, and Huiping Zhuang. L3a: Label-augmented analytic adaptation for multi-label class incremental learning. *arXiv preprint arXiv:2506.00816*, 2025b.

Haifeng Zhao, Yuguang Jin, and Leilei Ma. Dynamic prompt adjustment for multi-label class-incremental learning. In *International Conference on Brain Inspired Cognitive Systems*, pp. 52–61. Springer, 2024.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024.

A APPENDIX

This appendix holds the material referenced from the main text. Section A.1 documents how the per-eye labels are derived and how the eight classes co-occur, Section A.2 gives every metric for every method, and Sections A.4 to A.8 report five supporting studies that the main text mentions but has no room to show. It contains the reversed schedule, the cost of freezing a representation around the base task, the penalty-strength sweeps, the per-seed paired differences behind every comparison we state as a result, and the decomposition of the naive joint gap along timing and locus. Moreover, it includes an additional experiment comparing task size with task order A.3.

A.1 PER-EYE LABELS AND LABEL CO-OCCURRENCE

ODIR-5K distributes one eight-dimensional label vector per patient, together with a free-text diagnostic keyword string for each eye. We derive one label vector per eye from that eye's keyword string, for the reason given in Section 4. Keywords describing image quality rather than a condition are discarded: lens dust, optic disk photographically invisible, low image quality, image offset, anterior segment image, and no fundus image. Each remaining keyword then sets one class. We check the derivation by aggregating the per-eye labels back to the patient level and comparing against the labels that the dataset ships. A patient is positive for a disease if either eye is positive, and normal only if both eyes are negative, which is how ODIR-5K defines a normal patient.

The two labels of a single task are both positive for 21 images (0.3%): 8 for $\{O, C\}$, 8 for $\{G, M\}$, and 5 for $\{A, H\}$. Across tasks, the co-occurrence is higher and occurs in the common conditions: diabetic retinopathy (D) appears with other (O) in 266 images and with hypertensive retinopathy (H) in 67 images. The only exclusive label is normal (N); none of the 3,125 images marked that way carries any other positive label, which is a semantic property of that label rather than of the grouping by frequency. The full assignment covers distinct keyword strings and is given verbatim in the supplementary code, together with the normalization applied before matching; a string absent from the map raises an error rather than defaulting to a class, so no eye acquires a label by fallback.

A.2 COMPLETE RESULTS FOR EVERY METHOD

Table 3 ranks every method by final mAP and reports retention and plasticity. Table 4 reports the calibration of the same final models in the same row order. Between them, they hold most of the numbers quoted in the main text: the naive and joint bounds and the metric dependence of the gap (Experiment 1), the two frozen variants (Experiment 2), the regularizer comparison, and the plasticity column behind the claim that stability rather than plasticity is the binding constraint (Experiment 3), and the head-refit diagnostic together with the calibration of every method (Experiment 4). Plasticity is the score a class reaches on the task that introduces it, before any later interference. As a result, a method with high plasticity and high forgetting learns well and retains badly.

Table 3: Final performance of every method, ordered by mean average precision.

Method	mAP	AUC	CF1	OF1	forget	plast.
Joint (upper bound)	**0.561**±0.010	**0.839**±0.009	0.536±0.015	**0.596**±0.012	—	0.561±0.010
Joint, reversed order	0.562±0.007	0.839±0.007	**0.540**±0.008	0.595±0.023	—	0.562±0.007
Head-refit (diagnostic)	0.489±0.016	0.814±0.010	0.483±0.018	0.527±0.021	—	—
KRT (Dong et al., 2023)	0.408±0.024	0.756±0.005	0.422±0.011	0.476±0.013	0.023±0.004	0.412±0.021
Anchoring (λ_{anc}=1)	0.407±0.008	0.764±0.009	0.422±0.009	0.476±0.016	0.032±0.002	0.410±0.011
Anchoring + max-entropy	0.406±0.005	0.766±0.013	0.435±0.013	0.483±0.015	0.033±0.007	0.412±0.011
EWC + anchoring	0.405±0.021	0.758±0.003	0.429±0.006	0.473±0.029	0.024±0.017	0.406±0.004
SEP-IN	0.397±0.001	0.724±0.007	0.414±0.008	0.434±0.033	0.000±0.000	0.397±0.001
Naive, reversed order	0.397±0.030	0.738±0.009	0.405±0.028	0.509±0.029	0.085±0.035	0.410±0.001
SEP-BASE	0.358±0.020	0.711±0.011	0.373±0.007	0.391±0.023	0.000±0.000	0.358±0.020
EWC (λ=10)	0.306±0.040	0.711±0.023	0.345±0.021	0.441±0.032	0.171±0.041	0.419±0.020
Naive (lower bound)	0.278±0.022	0.712±0.019	0.326±0.026	0.448±0.018	0.201±0.039	0.412±0.011

Table 4: Calibration of the same final models, in the same order as Table 3

Method	FPR	CR−CP	OR−OP	ECE	entropy
Joint (upper bound)	**0.099**±0.006	**0.075**±0.017	**0.176**±0.018	0.086±0.003	0.052±0.003
Joint, reversed order	0.096±0.021	0.037±0.092	0.159±0.070	0.087±0.005	0.050±0.001
Head-refit (diagnostic)	0.155±0.012	0.197±0.035	0.299±0.023	**0.041**±0.003	0.173±0.001
KRT (Dong et al., 2023)	0.211±0.029	0.189±0.102	0.385±0.057	0.222±0.006	0.373±0.003
Anchoring (λ_{anc}=1)	0.215±0.020	0.221±0.035	0.397±0.034	0.235±0.008	0.335±0.008
Anchoring + max-entropy	0.217±0.022	0.246±0.071	0.412±0.042	0.238±0.003	0.333±0.003
EWC + anchoring	0.228±0.026	0.255±0.057	0.423±0.028	0.234±0.003	0.333±0.010
SEP-IN	0.271±0.023	0.240±0.051	0.459±0.013	0.369±0.002	0.384±0.010
Naive, reversed order	0.141±0.003	0.081±0.035	0.238±0.025	0.116±0.004	0.035±0.000
SEP-BASE	0.334±0.040	0.335±0.083	0.503±0.036	0.391±0.003	0.294±0.014
EWC (λ=10)	0.237±0.031	0.257±0.055	0.395±0.024	0.233±0.001	0.087±0.004
Naive (lower bound)	0.251±0.047	0.233±0.081	0.437±0.083	0.233±0.001	0.092±0.002

A.3 Experiment 5: task size against task order

Experiment 1 reports that reversing the schedule lifts naive fine-tuning from 0.278 to 0.397 mAP and more than halves its forgetting. That comparison moves two things at once. The four tasks are ordered by class frequency, so they also differ by more than an order of magnitude in training size, from 3,512 images for T_1=$\{N, D\}$ down to 304 for T_4=$\{A, H\}$. Reversing the schedule, therefore, changes both where a task sits and how much data arrives at that position. This experiment separates the two.

We compare three arms, each run under both schedules with the same three seeds, splits, and hyperparameters. The first is the pair of runs from Experiment 1, at the full training budget with the task sizes as they fall. The second holds the budget fixed at $1,216$ images while keeping the original size ratios ($814/240/92/70$), which isolates the effect of training on a quarter of the data. The third uses the same budget and cuts every task to 304 images, the size of the smallest task in the training split, thereby removing the imbalance. Undersampling is stratified by the label pattern within a task - all image types are reduced in the same proportion, so the class balance is preserved. Each task uses the same sample in both directions, so only the order in which the model encounters them changes. The joint bound is retrained on the union of the four reduced samples, so that each arm is assessed against a ceiling at its own budget: 0.561 mAP on the full training set and 0.493 on $1,216$ images.

Each arm makes a prediction. If what helps is where a task sits in the sequence, the reversal should retain its advantage once all tasks are the same size. If what helps is how much data arrives at each position, the advantage should disappear. Figure 4 shows the second. Reversing the schedule is worth $+0.120$ mAP at the full budget, $+0.034$ once the budget is matched but the sizes are left as they fall, and -0.006 once every task is cut to 304 images. Forgetting moves with it, from -0.116 to -0.014 to $+0.033$. The frequent-to-rare schedule is almost flat across the three arms, at 0.278, 0.263, and 0.257, so the reduction is not simply making every run worse. What it removes is the advantage of reversing.

The per-class numbers show where that advantage came from. At the full budget, the reversal improves all eight classes. The two largest gains are cataract, $+0.291$, and myopia, $+0.300$. Balanced, both change sign, to -0.087 and -0.079. These are the two classes that $\{N, D\}$ follows under the reversed schedule, and $\{N, D\}$ carries $3,512$ of the $4,900$ training images. Annotation is complete in this benchmark, so it is reasonable that in the last task, all eight labels are observed. Those images supervise cataract and myopia once more (largely as negatives), since across the benchmark only 16 of the images in $\{N, D\}$ are cataract-positive and 8 are myopia-positive. However, nothing is stored, and the reversed schedule ends in a rehearsal pass. Cutting that pass to 304 images reduces the gain.

This still does not settle two things. The matched-budget arm has already lost most of the effect before any balancing. Hence, the experiment does not separate the cost of training on a quarter of the data from the cost of removing the imbalance. The difference $+0.034$ against a spread of 0.027 is not a quantity worth splitting. What it does establish is the null result: with equal task sizes, neither mAP nor forgetting responds to the order. The order effect in Experiment 1 reflects the size

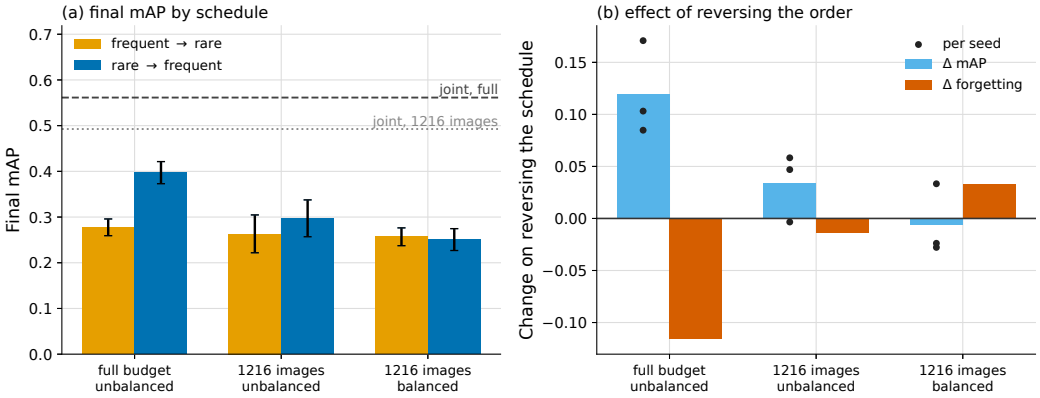

Figure 4: Experiment 5, separating the effect of the task order from the effect of task size. (a) Final mAP of naive sequential fine-tuning under each schedule. (b) The order effect itself, taken within each seed as rare-to-frequent minus frequent-to-rare, for mAP and for forgetting.

of the tasks in this dataset rather than their position, and the practical reading follows from this. A re-annotation programme whose last pass is also its largest is rehearsing old classes for free. One that ends on a rare condition is not, and should expect the forgetting that Experiment 1 measures on the frequent-to-rare schedule.

A.4 Effect of task order

Table 5 gives the per-class numbers behind the order effect reported in Experiment 1 - the task each class enters at under both schedules, and the naive final mAP, the forgetting, and the joint mAP for each. Joint training is unaffected by the order, so the table isolates the effect of the schedule. Cataract and myopia dominate the difference. When entered early and followed by three further tasks, they lose 0.552 and 0.440 mAP; when entered late, they lose 0.182 and 0.162. Normal and diabetic retinopathy enter last under the reversed schedule and so have no later task over which to decline. The rare classes move very little either way, which is why reversing the schedule changes the forgetting component while leaving the co-training component roughly where it was.

Table 5: Per-class effect of the task order

	C	M	N	O	D	H	A	G
T, frequent → rare	2	3	1	2	1	4	4	3
T, rare → frequent	3	2	4	3	4	1	1	2
naive mAP, freq → rare	0.301±0.115	0.403±0.157	0.465±0.033	0.234±0.038	0.446±0.070	0.151±0.052	0.093±0.026	0.128±0.035
naive mAP, rare → freq	0.592±0.098	0.703±0.100	0.627±0.031	0.355±0.025	0.480±0.041	0.185±0.011	0.103±0.041	0.133±0.054
forget, freq → rare	0.552±0.107	0.440±0.125	0.156±0.067	0.000	0.060±0.050	—	—	0.000
forget, rare → freq	0.182±0.121	0.162±0.128	—	0.000	—	0.038±0.050	0.054±0.042	0.076±0.020
joint mAP, freq → rare	0.894±0.007	0.895±0.019	0.694±0.020	0.450±0.024	0.546±0.039	0.241±0.096	0.411±0.131	0.359±0.062
joint mAP, rare → freq	0.882±0.003	0.895±0.012	0.694±0.014	0.468±0.038	0.542±0.043	0.242±0.096	0.415±0.197	0.358±0.064

A.5 Cost of stability

Figure 5 groups the final per-class mAP of the two frozen variants by the task that introduced each class. This is the evidence for the claim in Experiment 2 that adapting the backbone to the base task is paid for afterward. SEP-BASE is ahead on the T_1 classes it was fitted to and behind on the classes introduced later, most visibly myopia and hypertensive retinopathy. SEP-IN, which never adapts, is flatter across the schedule and ends 0.039 mAP ahead overall, on every seed.

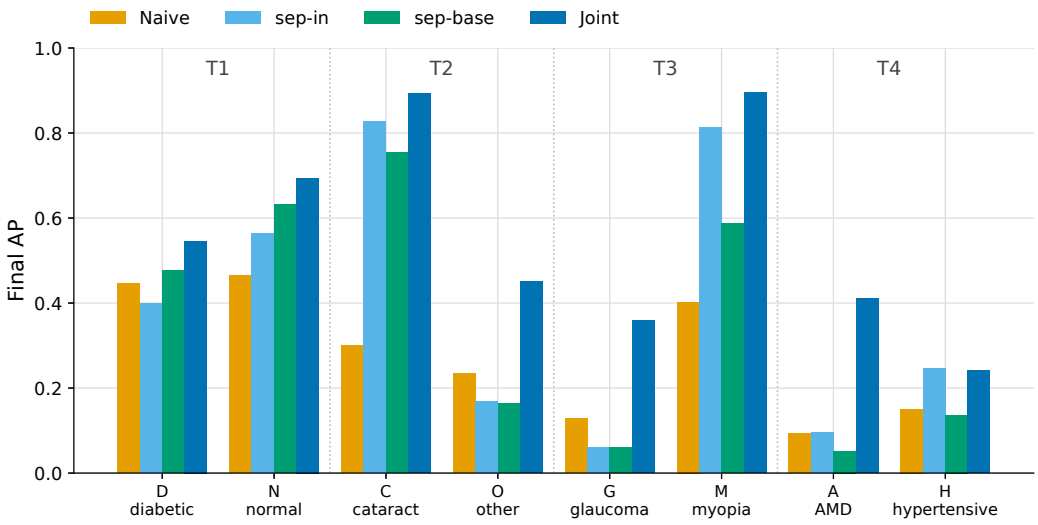

Figure 5: Per-class final mAP by introducing task (T_1 frequent to T_4 rare).

A.6 Penalty strength

Table 6 reports the sweeps behind the two settings quoted in Experiment 3. Anchoring is mildly sensitive, and trades retention against plasticity in the expected direction: forgetting falls from 0.046 to 0.023 across the range, and final mAP peaks at $\lambda_{\text{anc}}{=}1$. EWC is not sensitive at all. Three orders of magnitude of λ move its final mAP by 0.016 and leave forgetting between 0.171 and 0.186, which is why we read its failure as a property of the parameter-space penalty in this setting rather than as a badly chosen λ.

Table 6: Sensitivity of the two penalties to their strength, behind the settings quoted in Experiment 3.

Family	setting	mAP	forget	plast.
Anchoring	$\lambda_{\text{anc}}{=}0.5$	$0.393_{\pm 0.015}$	$0.046_{\pm 0.011}$	$0.411_{\pm 0.011}$
Anchoring	$\lambda_{\text{anc}}{=}1$	$0.407_{\pm 0.008}$	$0.032_{\pm 0.002}$	$0.410_{\pm 0.011}$
Anchoring	$\lambda_{\text{anc}}{=}2$	$0.400_{\pm 0.009}$	$0.023_{\pm 0.002}$	$0.404_{\pm 0.010}$
EWC	$\lambda{=}10$	$0.306_{\pm 0.040}$	$0.171_{\pm 0.041}$	$0.419_{\pm 0.020}$
EWC	$\lambda{=}10^2$	$0.290_{\pm 0.026}$	$0.186_{\pm 0.022}$	$0.416_{\pm 0.017}$
EWC	$\lambda{=}10^3$	$0.299_{\pm 0.018}$	$0.182_{\pm 0.023}$	$0.423_{\pm 0.014}$

A.7 Per-seed paired differences

Table 7 reports the difference taken within each of three seeds and then averaged, together with how many of the three seeds agree with the sign of that mean. The comparisons above the rule hold on all three seeds and are the ones we state as results. Those below it do not: EWC against naive, KRT against anchoring, the reversed joint bound against the forward one, and both additions to anchoring all change sign between seeds. We report them as differences we cannot resolve at this sample size, rather than as null results.

A.8 Two decompositions of the naive-joint gap

Experiments 1 and 4 divide the same quantity along different axes. The naive-joint gap is split by *timing* in Experiment 1, into a forgetting component (a class declines after its own task) and a lost co-training component (a class never reaches the joint level, although it never declines), and by *locus* in Experiment 4, into a decision-layer component (removed by refitting the heads on a frozen backbone) and a representation component (the residual that refitting cannot remove). Writing a_c^{nv},

Table 7: Paired differences in final mAP for every comparison the paper makes. Each difference is taken within a single pair of runs, on the same patient split and hyperparameters, and then averaged over the repetitions. Sign holds counts how many of those repetitions agree with the sign of the mean.

Comparison	Δ mAP	sd	sign holds
Head-refit − naive	$+0.211$	0.034	3/3
Anchoring − naive	$+0.129$	0.028	3/3
Naive reversed − naive	$+0.120$	0.045	3/3
Joint − head-refit	$+0.073$	0.024	3/3
SEP-IN − SEP-BASE	$+0.039$	0.019	3/3
Anchoring − SEP-IN	$+0.010$	0.007	3/3
EWC (λ=10) − naive	$+0.029$	0.040	2/3
KRT − anchoring	$+0.001$	0.016	2/3
Joint reversed − joint	$+0.001$	0.010	1/3
Anchoring + max-entropy − anchoring	-0.001	0.005	2/3
EWC + anchoring − anchoring	-0.002	0.013	2/3

a_c^{rf}, and a_c^{jt} for the final AP of class c under naive fine-tuning, the head refit, and joint training, and F_c for its forgetting as defined in Section 4,

$$\underbrace{a_c^{\mathrm{jt}} - a_c^{\mathrm{nv}}}_{\mathrm{gap}_c} = \underbrace{F_c}_{\text{forgetting}} + \underbrace{(\mathrm{gap}_c - F_c)}_{\text{co-training}} = \underbrace{(a_c^{\mathrm{rf}} - a_c^{\mathrm{nv}})}_{\text{decision layer}} + \underbrace{(a_c^{\mathrm{jt}} - a_c^{\mathrm{rf}})}_{\text{representation}}.$$

Both identities hold per class, so the two partitions can be compared directly, and Table 8 reports them.

Table 8: The naive-joint gap is decomposed by timing, locus, class, and group. Entries are computed per class from Table 2 and averaged over the eight classes, so the group rows are contributions to the eight-class mean, and every column adds to the total. Totals may differ by 0.001 from the values quoted in the main text, which are obtained by subtracting rounded aggregates. Forgetting is identically zero for the two classes introduced by T_4, which have no later task over which to decline; they are excluded from the per-class rate in Table 2 but enter the gap decomposition as zero.

			by timing		by locus	
Class	T	gap	forget	co-train	decision	repr.
declined after their own task						
C cataract	2	0.593	0.552	0.041	0.528	0.065
M myopia	3	0.492	0.440	0.052	0.424	0.068
N normal	1	0.229	0.156	0.073	0.203	0.026
D diabetic	1	0.100	0.060	0.040	0.049	0.051
group		0.177	0.151	0.026	0.151	0.026
never declined						
A AMD	4	0.318	—	0.318	0.204	0.114
G glaucoma	3	0.231	0.000	0.231	0.142	0.089
O other	2	0.216	0.000	0.216	0.071	0.145
H hypertensive	4	0.090	—	0.090	0.066	0.024
group		0.107	0.000	0.107	0.060	0.047
all eight classes		0.284	0.151	0.133	0.211	0.073

The decision-layer component (0.211) exceeds the entire forgetting component (0.151). The excess, 0.060, is accounted for exactly by the four classes that never declined: their gaps are co-training by construction, yet refitting the heads on an unchanged backbone recovers 0.204, 0.142, 0.071 and

0.066 mAP of them. Refitting therefore reaches 45% of the co-training component that Experiment 1 identifies as the part that no buffer-free method recovers during training. On the other hand, two-thirds of the representation residual (0.047 of 0.073) is concentrated in the same four classes, which is the quantitative form of the observation in Experiment 4 that the residual is visible on the rare co-occurring classes.

For the four classes that declined, the forgetting component and the decision-layer component agree to 0.0005 mAP (0.1510 against 0.1505). Although they are computed from independent measurements: one per task, the other from a separate refit run. Refitting restores these classes to approximately their peak values on their own task. The agreement is a group-level property and conceals compensating per-class deviations: the refit lands 0.024, 0.016, and 0.011 mAP below the peak for cataract, myopia, and diabetic retinopathy, and 0.047 above it for normal, which is to say that refitting gives the normal class a score it never attained during sequential training.

A decomposition is informative only if its components react differently to a manipulation aimed at one of them. Reversing the schedule targets forgetting, and Table 9 shows that it does: forgetting falls by 58% while co-training falls by 24%. The residual movement in the co-training component is expected, since reversing the order also changes which classes end the sequence, thereby shifting the training subpopulation.

Table 9: The timing decomposition under both schedules. Joint training is unaffected by the order (0.562 against 0.561 mAP), so the two rows differ only in the position of the tasks. Forgetting is averaged over all eight classes, as in Table 8; the six-class averages are the values reported in Table 3.

Schedule	gap to joint	forgetting	co-training
frequent $\rightarrow$ rare	0.284	0.151	0.133
rare $\rightarrow$ frequent	0.165	0.064	0.101
relative change		-58%	-24%

The head refit is a diagnostic, not a method. It uses the complete label matrix from the training split, which no method operating under the constraints of this study can access. Therefore, Table 8 localizes where the loss resides, and does not claim that a deployable buffer-free method can reach it. All entries are means over the three seeds and inherit the per-class dispersion reported in Table 2, which is largest for the rare classes.