Kaizen: Practical Self-supervised Continual Learning with Continual Fine-tuning

Chi Ian Tang [1,2] Lorena Qendro [1] Dimitrios Spathis [1] Cecilia Mascolo [2] Fahim Kawsar [1] Akhil Mathur [1]

[1]Nokia Bell Labs, UK [2]University of Cambridge, UK

WACV 2024, JAN 4-8, WAIKOLOA HAWAII

Overview

Kaizen strikes an optimal balance between **supervised** and **unsupervised methods**, grounded in more **realistic data assumptions** and enhanced **usability over time**.

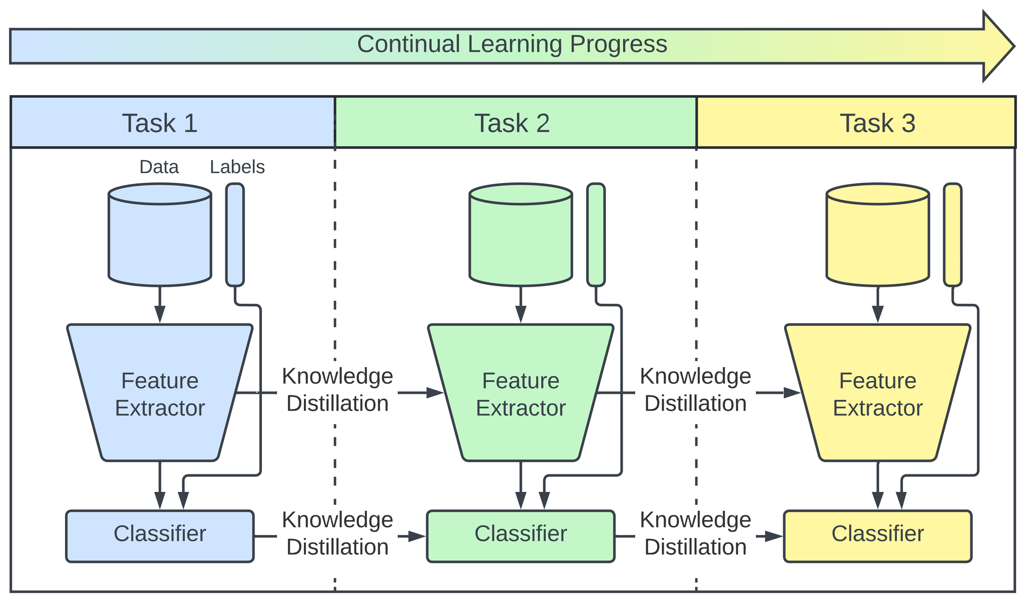

Figure: Representation of Kaizen providing continual fine-tuning.

- Kaizen ensures **reliable usability of the classifier**, while it continuously adapts to new tasks.
- It leverages labeled data as available, **combining self-supervised learning with unlabeled data** for enhanced generalization and reduced dependence on labels.
- Kaizen effectively harmonizes **learning new tasks** with the **retention of existing knowledge**.

Method

- Kaizen proposes a **joint loss function** that balances learning objectives, allowing models to **learn** from new data while **retaining knowledge** from previous tasks:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FE}}^{\mathrm{KD}} + \mathcal{L}_{\mathrm{C}}^{\mathrm{KD}} + \mathcal{L}_{\mathrm{C}}^{\mathrm{CT}} + \mathcal{L}_{\mathrm{FE}}^{\mathrm{CT}}$$

- The feature extractors are trained through SSL alongside knowledge distillation, while the classifiers are trained on both **unlabelled and labelled data** through **knowledge distillation and fine-tuning**.

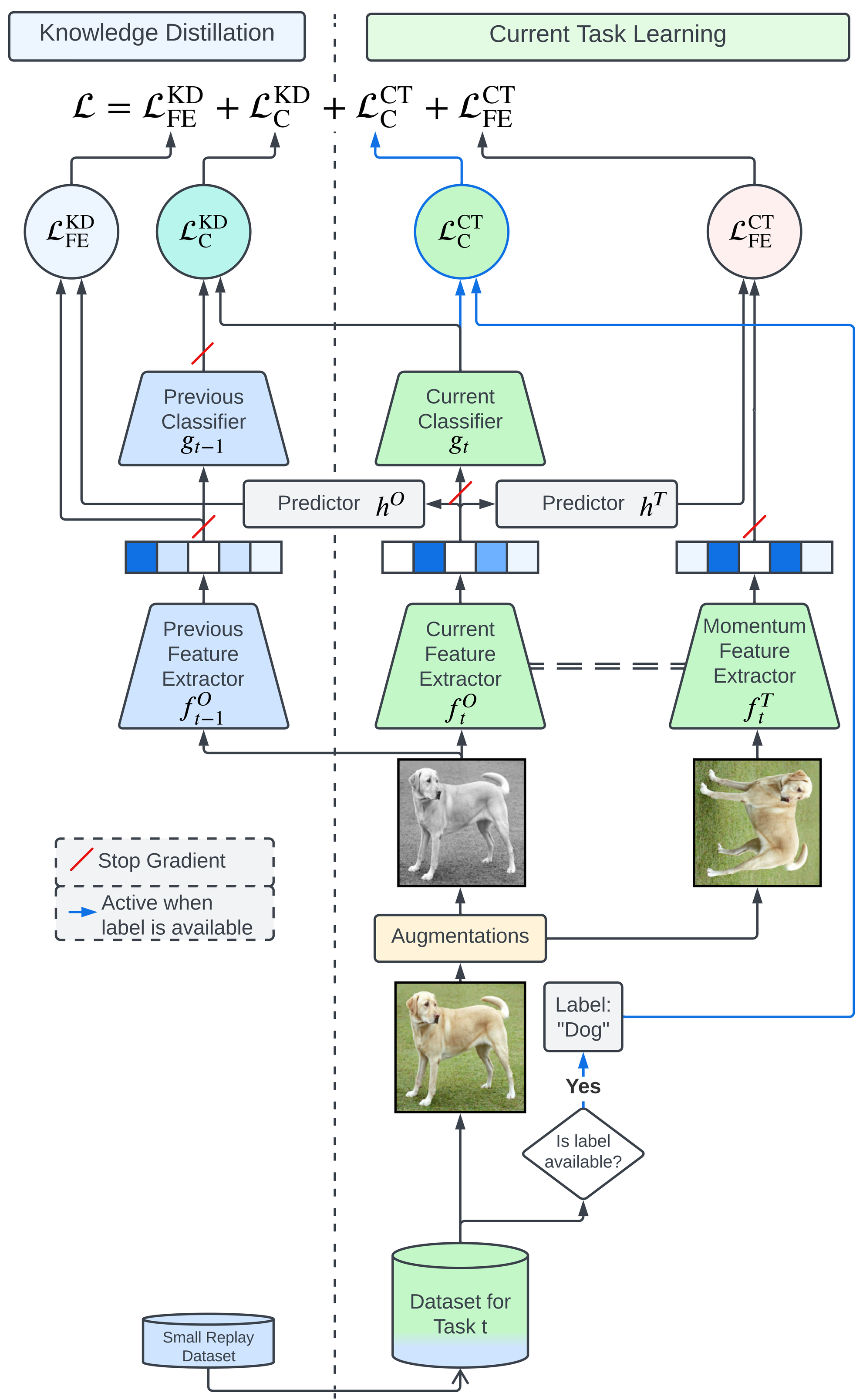

Figure: Overview of the Kaizen framework.

Kaizen is made of the following components that **tackle catastrophic forgetting and label scarcity**:

1. Contrastive self-supervised learning module for the feature extractor
2. Knowledge distillation for the feature extractor
3. Supervised learning module for the classifier
4. Knowledge distillation for the classifier
5. Memory replay: a subset of samples from previous tasks are replayed during training.

Evaluation

- **Datasets.** CIFAR-100 and ImageNet100, randomly divided into equal-class tasks for the experiments.
- **SSL Methods.** SSL backbones like SimCLR, MoCoV2+, BYOL, and VICReg to assess its generalizability.
- **Baselines.** SOTA CSSL pipeline - CaSSLe [1] and a no distillation setup.

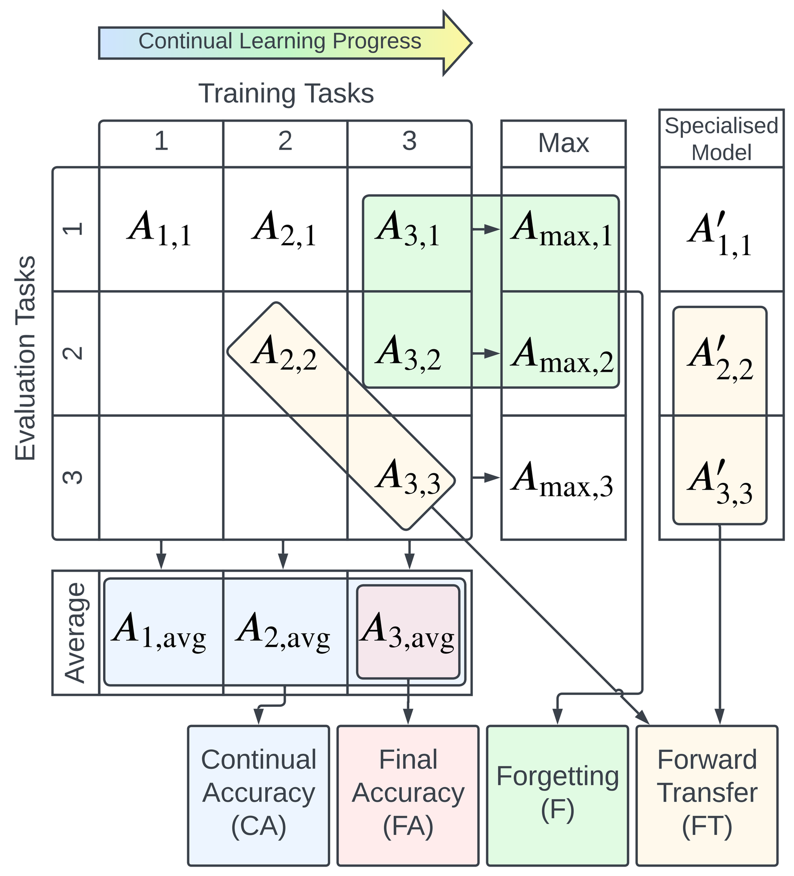

Figure: Calculation of the evaluation metrics.

Metrics:

- Continual Accuracy. Accuracy throughout the learning process.
- Final Accuracy. Accuracy at the final learning step.
- Forgetting. Performance lost over time.
- Forward Transfer. Specialised vs continual learning model.

Performance overview

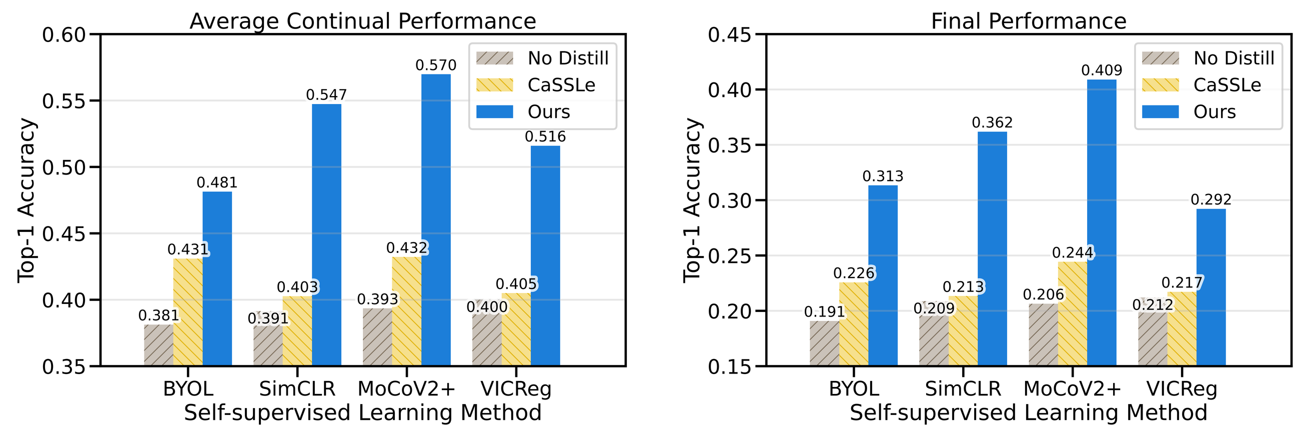

Figure: Performance comparison on CIFAR100. The left figure shows the average performance across the entire continual learning process, while the right figure shows the performance in the final evaluation.

- Kaizen is compatible with different SSL methods.
- **Outperforms SOTA** method CaSSLe by up to **13.8%** in Continual Accuracy.

Per-task performance breakdown

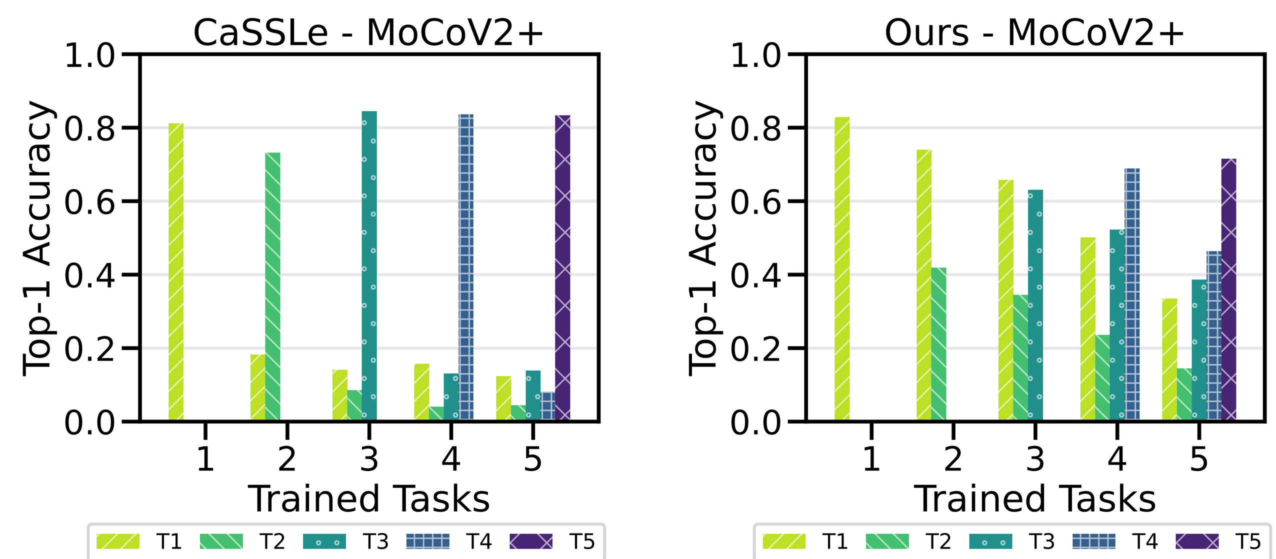

Figure: Detailed breakdown of performance over tasks on CIFAR-100. Fine-grained accuracy for every additional task across Kaizen and CaSSLe, with a fixed SSL backbone.

- **Kaizen exhibits a more refined degradation of performance on previous tasks over time**, by forgetting acquired knowledge in a more controlled and graceful manner.
- **SOTA** ignores knowledge distillation for the classifier and **suffers a significant one-step drop in performance on previous tasks.**

Performance variation over time

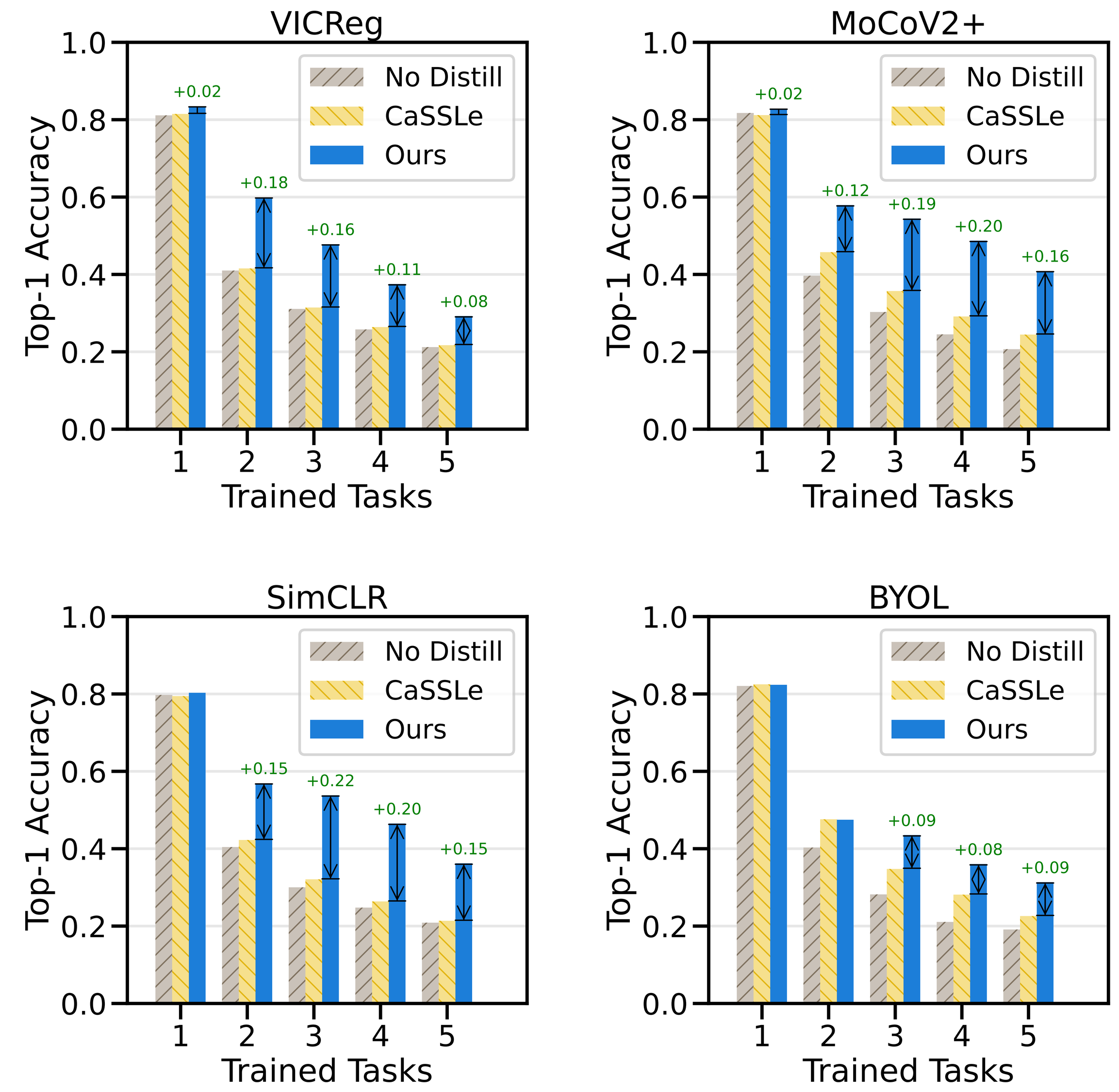

Figure: Average performance over 5 tasks on CIFAR-100. Each group of bars represents performance at each continual learning step.

- The **performance naturally decreases**, because the task becomes more and **more difficult over time with more classes.**
- Our method enables **more gradual forgetting over time**, which results in **higher final performance across time.**

Ablation on replay dataset size

- **Replay** is a crucial component in **combating catastrophic forgetting** for methods that do not rely on task labels.

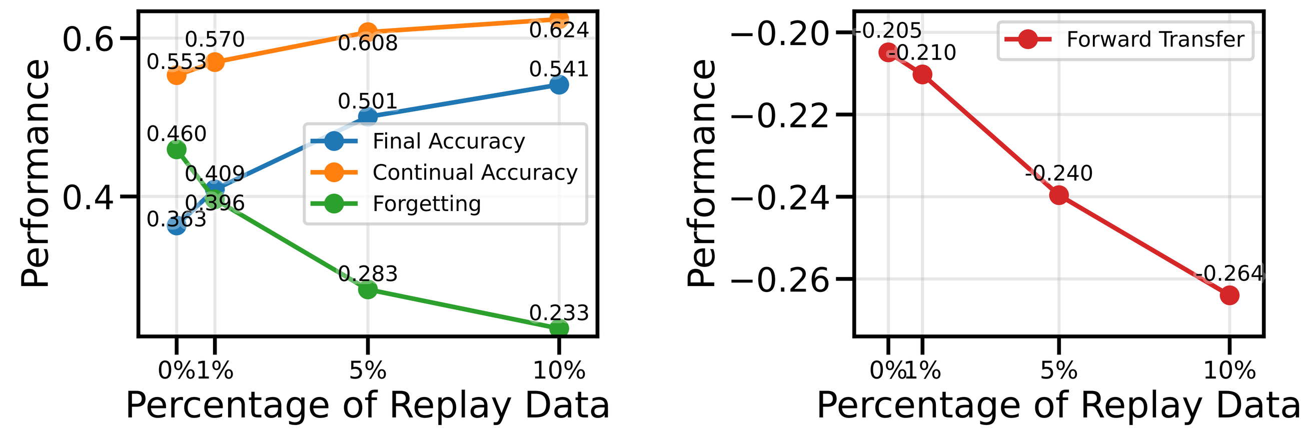

Figure: Performance of Kaizen on CIFAR100 with MoCoV2+ backbone with varying amounts of replay data.

- More replay data improves overall performance and reduces forgetting, though at the expense of forward transfer.
- **Kaizen performs well even with zero replay**, meeting stringent data conditions where replay is impractical.

Conclusions

- Kaizen improves existing continual learning methods, enabling **ongoing classifier training and flexible deployment at any stage.**
- Its learning objectives are strategically designed to ensure a **balanced training of the feature extractor and classifier.**
- Rigorous testing with a broad set of evaluation metrics shows that **Kaizen excels in balancing knowledge retention with new learning**, outperforming current methods.

References and Links

[1] Enrico Fini, Victor G Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal.
Self-supervised models are continual learners.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621--9630, 2022.

Paper

Code