# FLUID: A Unified Evaluation Framework for Flexible Sequential Data

Matthew Wallingford[†], Aditya Kusupati[*†], Keivan Alizadeh-Vahid[*†],
Aaron Walsman[†], Aniruddha Kembhavi[†‡] and Ali Farhadi[†]
[†]University of Washington, [‡]Allen Institute for Artificial Intelligence
https://raivn.cs.washington.edu/projects/FLUID/

## Abstract

*Modern ML methods excel when training data is IID, large-scale, and well labeled. Learning in less ideal conditions remains an open challenge. The sub-fields of few-shot, continual, transfer, and representation learning have made substantial strides in learning under adverse conditions; each affording distinct advantages through methods and insights. These methods address different challenges such as data arriving sequentially or scarce training examples, however often the difficult conditions an ML system will face over its lifetime cannot be anticipated prior to deployment. Therefore, general ML systems which can handle the many challenges of learning in practical settings are needed. To foster research towards the goal of general ML methods, we introduce a new unified evaluation framework –* FLUID *(Flexible Sequential Data).* FLUID *integrates the objectives of few-shot, continual, transfer, and representation learning while enabling comparison and integration of techniques across these subfields. In* FLUID*, a learner faces a stream of data and must make sequential predictions while choosing how to update itself, adapt quickly to novel classes, and deal with changing data distributions; while accounting for the total amount of compute. We conduct experiments on a broad set of methods which shed new insight on the advantages and limitations of current solutions and indicate new research problems to solve. As a starting point towards more general methods, we present two new baselines which outperform other evaluated methods on* FLUID*.*

## 1. Introduction

Modern ML methods have demonstrated remarkable capabilities, particularly in settings with large-scale labeled training data drawn IID. However, in practice the learning conditions are often not so ideal. Consider a general recognition system, a key component in many computer vision applications. One would expect such a system to learn from a changing data distribution, recognize classes with few and many examples, revise the set of known classes as new ones are seen, and update itself over time using new data.

[*]Equal contribution

Various subfields such as few-shot, continual, transfer, and representation learning have made substantial progress on the challenges associated with learning in non-ideal settings. The methods from these fields excel when the deployment conditions can be anticipated and align with specific scenarios. For example, few-shot methods perform well when the number of new classes and examples per class are known in advance. Similarly, continual learning techniques improve performance when new data arrives in fixed-size batches at predictable intervals. However, in many applications the exact conditions cannot be known a priori and are likely to change over time. Data may come at irregular intervals, the volume of supervision for each class may change, or new classes might be encountered. This calls for general methods that can handle a plethora of scenarios during deployment.

To make progress towards such general methods, new evaluations which reflect the key aspects of learning in practical settings are essential. But, *what are these aspects?* We posit the following as some of the necessary elements: (1) *Sequential Data* - In many application domains the data streams in. ML methods must be capable of learning from new data and changing distributions. (2) *X-shot* - Data often has a different number of examples for each class (few for some and many for others). Current evaluations assume prior knowledge of which data regime (few-shot, many-shot, etc.) new classes will be from, but often this cannot be known in advance. (3) *Flexible Training Phases* – Practical scenarios rarely delineate when and how a system should train. ML systems should be capable of making decisions such as when to train based on incoming data, which data to train with, and whether to update all parameters or just the classifier. (4) *Compute Aware* – Real-world systems often have computational constraints not only for inference but also for training. ML systems should account for the total compute used throughout their lifetime. (5) *Open-world* - As new data is encountered the set of known classes may change over time. Learning in practical settings often entails recognizing known classes while detecting when data comes from new classes.

With these elements in consideration we introduce FLUID (**F**lexib**l**e seq**u**enti**i**al **d**ata), a unified evaluation framework.

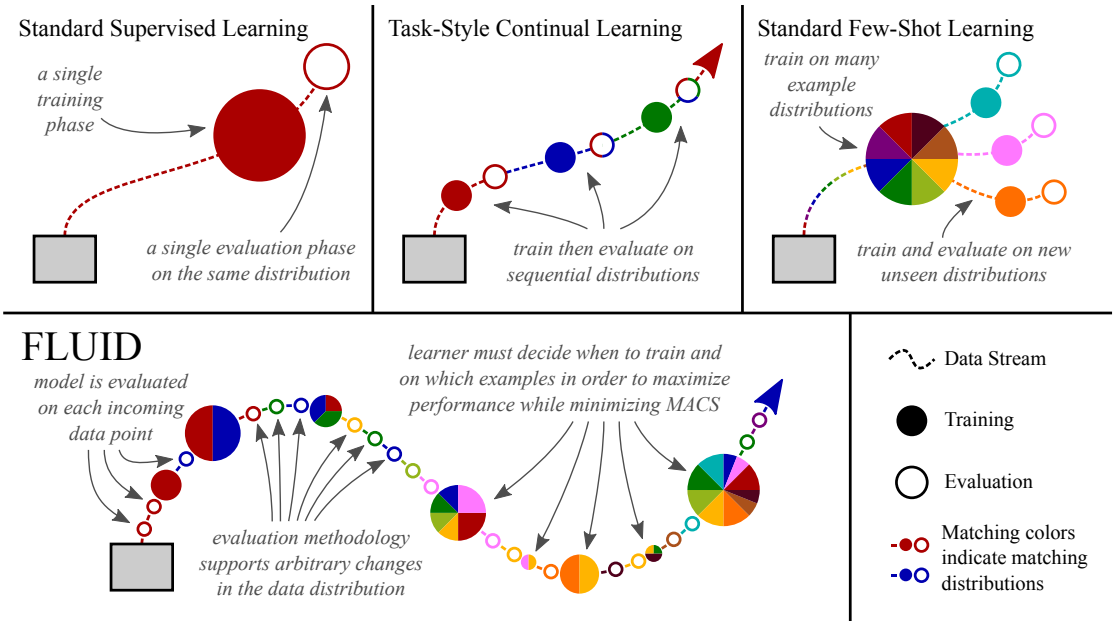



Figure 1. Comparison of supervised (top-left), continual (top-middle), and few-shot learning (top-right) with FLUID (bottom). The learner (grey box) accumulates data (dotted path), trains on given data (filled nodes), then is evaluated (empty nodes). The size of the node indicates the scale of the training or evaluation. Each color represents a different set of classes.

FLUID integrates the objectives of few-shot, continual, transfer, and representation learning into a simple and realizable formulation along with a benchmarkable implementation. In FLUID, a learner is deployed on a stream of data from an unknown distribution and must classify incoming samples one at a time while deciding when and how to update based on newly received data.

We conduct extensive experiments with FLUID on a broad set of methods across subfields. The experimental results quantitatively demonstrate the current limitations and capabilities of various ML approaches. For example, we find that canonical few-shot methods do not scale well to scenarios with more classes and varying number of examples. Similarly, we observe that continual learning methods often hinder performance when large-scale pretraining is readily available. Finally, we briefly explore the emergent problem of finding update strategies for deciding when and how to train efficient on incoming data. The framework, data and models will be open-sourced.

**We make the following contributions:**

1. A new evaluation framework, FLUID, which unifies the objectives of few-shot, continual, transfer, and self-supervised learning into a simple evaluation that enables comparison and integration of methods across related subfields and presents new research challenges.
2. New insights from experiments with FLUID like higher capacity networks generalize better to novel classes, representative self-supervised methods behave differently than standard pretraining on few-shot classes, and large-scale pretraining mitigates catastrophic forgetting.
3. Two simple baselines, Exemplar Tuning & MDT, which outperform evaluated methods in FLUID while matching performance in supervised and few-shot settings.

## 2. FLUID Related Work

We discuss the key aspects of FLUID in the context of other works and evaluations. Logistical details of FLUID are provided in section 3. Lastly, We outline the existing frameworks along with the supported properties in Table 1

**Sequential Data and Continual Learning** New data is an inevitable consequence of our dynamic world and learning over time is a long-standing challenge [53]. In recent years, continual learning (CL) has made notable progress on the problem of learning in a sequential fashion [1, 2, 25, 33, 46, 47]. Several setups have been proposed in order to evaluate systems' abilities to learn continuously and primarily focus on *catastrophic forgetting*, a phenomenon where models drastically lose accuracy on old tasks when trained on new tasks. The typical CL setup sequentially presents data from each task then evaluates on current and previous tasks [25, 33, 46]. Recent variants have proposed a task-free setting where the data distribution changes without the model's knowledge [18, 19, 47, 52, 63].

There are two assumptions in CL setups which we remove in FLUID. The first assumption is that data will be received in large batches with ample data for every class in the task.

Table 1. We categorize existing evaluations frameworks aimed at learning in practical settings. ✓: presence; ✗: absence & –: not applicable.

| Framework \ Property | Open-world | Sequential | Variable Batch-size | Few-shot | Many-shot | Compute Aware | Memory Constrained | Flexible Training | Non-IID |
|---|---|---|---|---|---|---|---|---|---|
| Representation Learning | ✗ | ✗ | – | ✗ | ✓ | – | – | – | ✓ |
| Transfer Learning | ✗ | ✗ | – | ✗ | ✓ | – | – | – | ✓ |
| Task-based Continual Learning | ✗ | ✓ | ✗ | ✗ | ✓ | – | ✓ | ✗ | ✓ |
| Task-free Continual Learning | ✗ | ✓ | ✗ | ✗ | ✓ | – | ✓ | ✗ | ✓ |
| Few-shot Learning | ✗ | ✗ | – | ✓ | ✗ | – | – | – | ✓ |
| Generalized Few-shot Learning | ✗ | ✗ | – | ✓ | ✓ | – | – | – | ✓ |
| Streaming Perception | ✗ | ✓ | – | – | – | ✓ | – | – | ✗ |
| Open Long-Tailed Recognition | ✓ | ✗ | – | ✓ | ✓ | – | – | – | ✗ |
| Data Stream Classification | – | ✓ | ✓ | – | ✓ | – | ✓ | – | ✓ |
| Test Time Training | ✗ | ✓ | ✓ | – | – | – | – | ✗ | ✓ |
| FLUID **(Ours)** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | – | ✓ | ✓ |

This circumvents a fundamental challenge of sequential data which is learning new classes from only a few examples. Consider the common scenario in which a learner encounters an instance from a novel class. The system must determine that it belongs to a new class with no previous examples (zero-shot learning and out-of-distribution detection). The next time an instance from the category appears, the system must be capable of one-shot learning, and so forth. In other words, few-shot learning is a natural consequence of learning from sequential data. Second assumption is that the training and testing phases will be delineated to the system. Deciding when to train and which data to train on is vital and an intrinsic challenge of learning continuously.

Some continual learning formulations include a memory cache to store images, typically between 0 - 1000, from previous tasks. We argue that setting a specific memory constraint, particularly this small, is too constraining. Methods should account for memory, but the FLUID framework does not explicitly restrict memory during streaming. Note that for a fair comparison to CL methods we use no memory caching for the Nearest Class Mean (NCM) baseline and complete caching for CL methods.

Lastly, data stream classification [4, 14, 50, 60] has worked on the problem of learning from sequential data. This line of work primarily focuses on traditional classification and not image recognition. At a high level FLUID has similar goal as data stream classification. FLUID differs in its implementation which importantly integrates the preexisting fields of few-shot, transfer, representation, and continual learning and provides a concrete benchmark for sequential image recognition.

**Few-shot and X-shot Learning** Learning from few examples for some classes is an inherent aspect of the real-world. Learning from large, uniform datasets [34, 48] has been the primary focus of supervised learning while few-shot learning has gained traction as a subfield [17, 38, 45, 51].

While few-shot learning is a step towards more generally applicable ML methods, the framework has assumptions that are unlikely to hold in practical settings. The experimental setup for few-shot is typically the $n$-shot $k$-way evaluation. Models are trained on base classes during *meta-training* and then tested on novel classes during *meta-testing*. The $n$-shot $k$-way experimental setup is limited in two respects. $n$-shot $k$-way assumes that a model will always be given exactly $n$ examples for $k$ classes at test time which is unrealistic. Second, most works only evaluate 5-way scenarios with 1, 5, and 10 shots. Realistic settings often have a mix of classes from both the high and low data regime. Expecting all classes to have less than 10 samples is not a reasonable assumption. Recently, general variants of few-shot learning have been proposed [6, 11] and are still far from being practical.

FLUID naturally integrates the few-shot problem into its framework by sequentially presenting data from a long tail distribution and evaluates systems across a spectrum of shots and ways. Our experimental results on representative few-shot methods indicate that they are overly tuned to the specific conditions of the few-shot evaluation which validates the need for a more general framework such as FLUID.

**Flexible Training Phases** Current experimental setups dictate when models will be trained and tested. Ideally, an ML system should be capable of knowing when to train itself, what data to train on, and what to optimize for [9]. By removing the assumption that training and testing phases are fixed and known in advance, FLUID provides a benchmark for tackling the relatively unexplored problem of learning when to train.

**Compute Aware** ML systems capable of adapting to their environment over time must account for the computational costs of their learning strategies as well as of inference. Prabhu et al. [40] showed that current CL frameworks do not measure total compute and therefore a naive but compute-hungry strategy can drastically outperform state of the art methods. Previous works have primarily focused on efficient inference [23, 28, 29, 35, 44] and some on training costs [12]. In FLUID we measure the total compute for both learning and inference over the sequence.

**Open-world** Practical scenarios entail inferring in an open world - where the classes and number of classes are

unknown to the learner. Few-shot, continual, and traditional supervised learning setups assume that test samples can only be from training classes. Previous works explored static open-world recognition [3, 26, 36, 43, 57] and the related problem of out-of-distribution detection [22, 32, 37]. FLUID presents a natural integration of sequential and open-world learning where the learner must identify new classes and update its known set of classes throughout the stream.

## 3. FLUID Evaluation Details

FLUID evaluation is designed to be simple and general while integrating the key aspects outlined in section 2.

**Formulation** Let learning system $S$ be composed of a model, $f_\theta : \mathbf{x} \mapsto \mathbf{y}$, and update strategy, $U : f_\theta \times \bigcup_{t=1}^{T}(x_t, y_t) \mapsto f_{\theta'}$, where $\bigcup_{t=1}^{T}(x_t, y_t)$ is the training data collected up to time $T$. Model, $f_\theta$, may be initialized using pretraining data $D = \{x_i, y_i\}_{i=1}^{n}$.

At each new time step, $t+1$, the model is given a sample, $x_t$, and provides a class label, $f_\theta(x_{t+1}) \in \{1...K+1\}$, for $K$ known classes. In other words, the sample may belong to one of $K$ previously seen classes, or a new class. The model output is evaluated with respect to the true label, $\mathbb{1}\{f_\theta(x_{t+1}) = y_{t+1}\}$, and $(x_{t+1}, y_{t+1})$ is added to the training set. If $y_{t+1}$ is from a new class, the set of known classes is updated accordingly. Next the model, $f_\theta$, may be updated according to $U$ using all previously observed data. This process is repeated for some total number of time steps.

Systems are evaluated on a suite of metrics including the overall and mean class accuracy throughout the stream along with the total compute required for updates and inference. Lastly, Algorithm 1 in Appendix A provides a psuedo code for the implementation of the FLUID framework.

**Data** In this paper, we evaluate methods with FLUID using a subset of ImageNet-22K [10]. Traditionally, few-shot learning used datasets like Omniglot [30] & MiniImagenet [58] and continual learning focused on MNIST [31] & CIFAR [27]. Some recent continual learning works have used Split-ImageNet [61]. The aforementioned datasets are mostly small-scale and have very few classes. We evaluate on the ImageNet-22K dataset to present new challenges to existing models. Recently, the INaturalist [56, 62] and LVIS [16] datasets have advocated for heavy-tailed distributions. We follow suit and draw our sequences from a heavy-tailed distribution.

The dataset consists of a pretraining dataset and 5 different sequences of images for streaming (3 test and 2 validation sequences). For pretraining we use the standard ImageNet-1K [48]. This allows us to leverage existing models built by the community as pre-trained checkpoints. Sequence images come from ImageNet-22K after removing ImageNet-1K's images. Each test sequence contains images from 1000 different classes, 750 of which do not appear in ImageNet-1K. We refer to the overlapping 250 classes as Pretrain classes and the remaining 750 as Novel classes. Each sequence is constructed by randomly sampling images from a heavy-tailed distribution of these 1000 classes. Each sequence contains $\sim 90000$ samples, where head classes contain $> 50$ and tail classes contain $\leq 50$ samples. The sequence allows us to study how methods perform on combinations of pretrain vs novel, and head vs tail classes. In Table 3, we show results obtained for sequence 5, and the Appendix I shows results across all test sequences. More comprehensive statistics on the data and sequences are in Appendix B.

**Pretraining** Supervised pretraining [21] on large annotated datasets like ImageNet facilitates the transfer of learnt representations to help data-scarce downstream tasks. Unsupervised learning methods like autoencoders [55] and more recent self-supervision methods [15, 24, 41] like Momentum Contrast (MoCo) [20] and SimCLR [7] have begun to produce representations as rich as that of supervised learning and achieve similar accuracy on various downstream tasks.

Before the sequential phase, we pretrain our model on ImageNet-1K. In our experiments, we compare how different pretraining strategies (contrastive learning, meta-training, & supervised training) generalize under more open and challenging conditions. We find new insights such as contrastive representations perform significantly worse on few-shot classes compared to supervised counterparts in the FLUID evaluation and meta-training causes larger networks to lose performance in a way that supervised training does not.

**Evaluation metrics** Table 2 defines the evaluation metrics in FLUID to gauge the performance of the algorithms.

Table 2. The evaluation metrics used in the FLUID framework to capture the performance and capabilities of various algorithms.

| Metric | Description |
|---|---|
| Overall Accuracy | Accuracy over the sequence. |
| Mean Per-Class Accuracy | Accuracy averaged over all classes in the sequence. |
| Total Compute | MAC operations for all compute over the sequence. |
| Unseen Class Detection | AUROC for the detection of OOD samples. |
| Cross-Sectional Accuracies | Classes in the sequence belong fall into 4 subcategories:<br>1) *Pretraining-Head*: $> 50$ samples & in pretraining.<br>2) *Pretraining-Tail*: $\leq 50$ samples & in pretraining.<br>3) *Novel-Head*: $> 50$ samples & not in pretraining.<br>4) *Novel-Tail*: $\leq 50$ samples & not in pretraining. |

## 4. Baselines and Methods

We summarize the baselines, other methods, and our proposed baselines, Exemplar Tuning and MDT. Additional details about the methods and implementation can be found in Appendix D and Appendix E respectively.

**Standard Training and Fine-Tuning** We evaluate stan-

dard model training (update all parameters in the network) and fine-tuning (update only the final linear classifier) with offline batch training. We ablate over the number of layers trained during fine-tuning in Appendix F.

**Nearest Class Mean (NCM)** Recently, multiple works [54, 59] have found that Nearest Class Mean (NCM) is comparable to state-of-the-art few-shot methods [38, 51]. NCM in the context of deep learning performs a 1-nearest neighbor search in feature space with the centroid of each class as a neighbor. We pretrain a neural network with a linear classifier using softmax cross-entropy loss, then freeze the parameters to obtain features.

**Few-shot Methods** We evaluate four representative methods: MAML [13], Prototypical Networks (PTN) [49], Weight Imprinting [42] & Meta-Baseline [8].

PTN trains a deep feature embedding using 1-nearest neighbor with class centroids and soft nearest neighbor loss. Parameters are trained with meta-training and backprop.

MAML is a gradient-based approach which uses second-order optimization to learn parameters that can be quickly fine-tuned and adapt to a given task. We tailor MAML to FLUID by pretraining the model according to the objective in Appendix D and then fine-tune during the sequential phase.

Weight Imprinting initializes the weights of a cosine classifier as the class centroids, then fine-tunes with a learnable temperature. For further analysis of Weight Imprinting and comparison to Exemplar Tuning see Appendix L.

Meta-Baseline is the same as NCM in implementation except that a phase of meta-training is done after regular softmax cross-entropy pretraining.

**Continual Learning (CL) Methods** We evaluate Learning without Forgetting (LwF) [33] and Elastic Weight Consolidation (EWC) [25] to observe whether continual learning techniques can improve performance in FLUID.

LwF leverages knowledge distillation [5] to retain accuracy on previous training data without storing it. EWC enables CL in a supervised learning context by penalizing the total distance moved by the parameters from the optimal model of previous tasks weighted by the corresponding Fisher information. Unlike LwF, EWC requires stored data, typically the validation set, from the previous tasks. In FLUID, we use LwF and EWC to retain performance on pretrain classes. For further details see Appendix E.

**Out-of-Distribution (OOD) Methods** We evaluate two methods proposed by Hendrycks & Gimpel [22] (HG) and OLTR [36] along with our proposed OOD baseline. The HG baseline thresholds the maximum probability output of the softmax classifier to determine whether a sample is OOD.

We propose the baseline, **Minimum Distance Thresholding (MDT)**, which utilizes the minimum distance from the sample to all class representations, $c_i$. In the case of NCM the class representation is the class mean and for a linear layer it is the $i$th column vector. For distance function $d$ and a threshold $t$, a sample is out of distribution if: $\mathbf{I}\left(\min_i d\left(c_i, \mathbf{x}\right) < t\right)$. Other metric learning techniques have proposed using distance to detect out of distribution examples [32, 37]. MDT primarily differs from these works in that it can be used with a standard classification network and can be performed in a single forward pass with negligible extra compute. MDT outperforms the other evaluated OOD methods in FLUID.

**Exemplar Tuning (ET)** We present a new method/baseline that leverages the inductive biases of instance-based methods and parametric deep learning. The traditional classification layer is effective when given a large number of examples but performs poorly when only a few examples are present. On the other hand, NCM and other few-shot methods are accurate in the low data regime but do not significantly improve when more data is added. Exemplar Tuning (ET) synthesizes these methods in order to initialize class representations accurately when learning new classes and to have the capacity to improve when presented with more data. We formulate each class representation (classifier), $C_i$, and class probability as the following:

$$C_i = \frac{1}{n}\sum_{x \in D_i} \frac{f(x;\theta)}{\|f(x;\theta)\|} + \mathbf{r}_i; \;\; p(y=i|x) = \frac{e^{C_i \cdot f(x;\theta)}}{\sum_{i \neq j} e^{C_j \cdot f(x;\theta)}} \tag{1}$$

where $f(x;\theta)$ is a parametrized neural network, $\mathbf{r}_i$ is a learnable vector, $n$ is the number of class examples, and $D_i$ are all examples in class $i$. Note that $C_i$ is comparable in form to the $i$-th column vector in a linear classification layer.

In this formulation, the class centroid (the first term of $C_i$ in Eq 1) provides an accurate initialization from which the residual term $\mathbf{r}_i$ can continue to learn. Thus ET is accurate for classes with few examples (where deep parametric models are inaccurate) and continues to improve for classes with more examples (where few-shot methods are lacking). In our experiments, we update the centroid after each sample with little additional compute and batch train the residual vector with cross-entropy loss according to the same schedule as fine-tuning (see Appendix E for implementation details).

Note that we compare ET to initializing a cosine classifier with class centroids and fine-tuning (Weight Imprinting). We find that Exemplar Tuning outperforms Weight Imprinting and affords two significant advantages besides better accuracy. 1) ET has two frequencies of updates (fast instance-based and slow gradient-based) which allows the method to quickly adapt to distribution shifts while providing the capacity to improve over a long time horizon. 2) ET automatically balances between few-shot & many-shot performance, unlike Weight Imprinting which requires apriori knowledge of when to switch from centroid-based initialization to fine-tuning.

Table 3. Performance of the suite of methods (outlined in Section 4) across accuracy and compute metrics on sequence 5. We present several variants of accuracy - Overall, Mean-per-class as well as accuracy bucketed into 4 categories: Novel-Head, Novel-Tail, Pretrain-Head and Pretrain-Tail (Pretrain refers to classes present in the ImageNet-1K dataset). Sup. refers to Supervised and MoCo refers to He et al. [20]. The best technique on every metric is in **bold**. Some methods could leverage caching of representations for efficiency, so, both GMACs are reported. GMACs do not include pretraining compute costs. See Table 7 in Appendix H for results with ResNet50 backbone.

| Method | Pretrain Strategy | Novel - Head (>50) | Pretrain - Head (>50) | Novel - Tail (<50) | Pretrain - Tail (<50) | **Mean Per-Class** | **Overall** | GMACs↓ ($\times 10^6$) |
|---|---|---|---|---|---|---|---|---|
| Backbone - Conv-4 | | | | | | | | |
| (a) Prototypical Networks | Meta | 11.63 | 22.03 | 6.90 | 13.26 | 11.13 | 15.98 | **0.06** |
| (b) MAML | Meta | 2.86 | 2.02 | 0.15 | 0.10 | 1.10 | 3.64 | 0.06 / 2.20 |
| Backbone - ResNet18 | | | | | | | | |
| (c) Prototypical Networks | Meta | 8.64 | 16.98 | 6.79 | 12.74 | 9.50 | 11.14 | 0.15 |
| (d) Meta-Baseline | Sup./Meta | 40.47 | 67.03 | 27.53 | 53.87 | 40.23 | 47.62 | 0.16 / 5.73 |
| (e) Weight Imprinting | Sup. | 40.32 | 67.46 | 15.35 | 34.18 | 32.69 | 48.51 | 0.16 / 5.73 |
| (f) LwF | Sup. | 30.07 | 67.50 | 7.23 | 56.96 | 31.02 | 48.76 | 22.58 / 45.16 |
| (g) EWC | Sup. | 39.03 | 70.84 | 16.59 | 47.18 | 34.89 | 50.39 | 12.29 |
| (h) OLTR | Sup. | 40.83 | 40.00 | 17.27 | 13.85 | 27.77 | 45.06 | 0.16 / 6.39 |
| (i) Fine-tune | Sup. | 43.41 | **77.29** | 23.56 | **58.77** | 41.54 | 53.80 | 0.16 / 5.73 |
| (j) Standard Training | Sup. | 38.51 | 68.14 | 16.90 | 43.25 | 33.99 | 49.46 | 11.29 |
| (k) NCM | Sup. | 42.35 | 72.69 | **31.72** | 56.17 | 43.44 | 48.62 | **0.15** |
| (l) Exemplar Tuning **(Ours)** | Sup. | **48.85** | 75.70 | 27.93 | 45.73 | **43.61** | **58.16** | 0.16 / 5.73 |
| (m) Weight Imprinting | MoCo | 16.77 | 26.98 | 6.19 | 8.69 | 12.60 | 22.90 | 0.16 / 5.73 |
| (n) OLTR | MoCo | 34.60 | 33.74 | 13.38 | 9.38 | 22.68 | 39.92 | 0.16 / 6.39 |
| (o) Fine-tune | MoCo | 14.49 | 27.59 | 0.10 | 4.96 | 8.91 | 26.86 | 0.16 / 5.73 |
| (p) Standard Training | MoCo | 26.63 | 45.02 | 9.63 | 20.54 | 21.12 | 35.60 | 11.29 |
| (q) NCM | MoCo | 19.24 | 31.12 | 14.40 | 21.95 | 18.99 | 22.90 | 0.15 |
| (r) Exemplar Tuning **(Ours)** | MoCo | 31.50 | 46.21 | 12.90 | 21.10 | 24.36 | 39.61 | 0.16 / 5.73 |

## 5. Experiments and Analysis

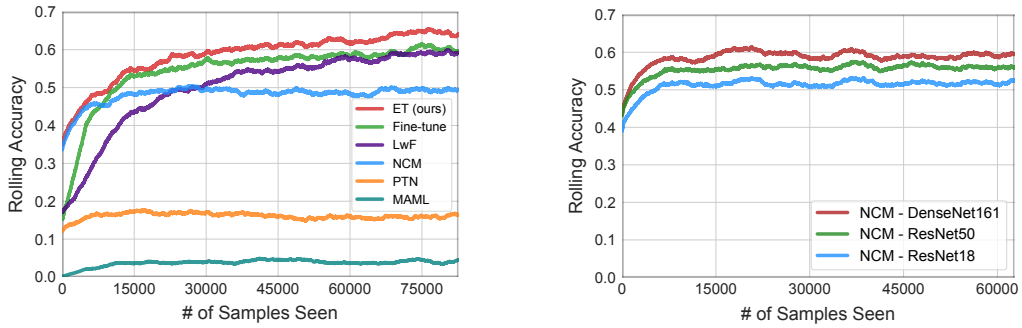
We evaluate representative methods from few-shot, continual, self-supervised learning, and out-of-distribution detection in the proposed FLUID framework. We present a broad set of practical observations and discuss novel findings that validate the need for FLUID and suggest future research directions. Table 3 displays a comprehensive set of metrics for the set of methods (outlined in Sec 4). Throughout this section, we will refer to rows of the table for specific analysis.

**Few-shot Analysis** We evaluate Prototypical Networks (PTN), Model Agnostic Meta-Learning (MAML), Weight Imprinting, and Meta-Baseline and compare to the baselines of Nearest Class Mean (NCM), fine-tuning, and standard training. We choose to evaluate each method for the following reasons:

- Prototypical Networks is a prominent metric learning method that is the basis for many other few-shot methods. Additionally PTN differs from NCM only in the use of meta-training so we can isolate the effect of meta-training.
- MAML is a canonical optimization-based meta-learning approach which many other methods extend from.
- Meta-Baseline is similar to NCM except that additional meta-training is performed after standard batch training. It is a simple method that is competitive with state of the art in few-shot learning.
- Weight Imprinting is a simple combination of NCM and fine-tuning and does not use meta-training. This makes for a natural comparison to both NCM and fine-tuning individually along with our proposed method, ET.

We find that PTN and MAML (Table 3-a,b,c and Figure 2-a) do not perform well on FLUID with over 30% lower overall accuracy than the NCM baseline (Table 3-k). One could argue that PTN and MAML may scale to the larger setting by increasing model capacity. However, few-shot works indicate that deeper and overparameterized networks decrease performance [13, 38, 49, 51]. We verify this observation, noting that the 4-layer convnet PTN (Table 3-a) outperforms the ResNet18 PTN in overall and novel class accuracy.

The prevailing thought in few-shot literature has been that smaller networks overfit less to base classes, and therefore methods use shallow networks or develop techniques to constrain deeper ones. We find evidence to the contrary, that deeper networks generalize better to novel classes with standard batch sampling (see Figure 2-b). Given that NCM and PTN differ only in the use of meta-training, we conclude that meta-training must be responsible for the lack of generalization in deeper networks. This evidence is further reinforced by the fact that meta-baseline performs worse than NCM




(a) (b)

Figure 2. (a) Compares the rolling accuracy of various methods over the stream of data. Exemplar Tuning performs the best at all stages of the stream. (b) Compares the accuracy of NCM on novel classes across network architectures. Contrary to prevailing thought, we find that deeper networks generalize better to novel few-shot classes.

with the inclusion of meta-training (Table 3-d,k). Although meta-training has merits for smaller data sets, in its current form it does not scale well to more data and classes.

While the few-shot benchmarks and methods have their merits and advantages, these observations validate the need for more general and challenging evaluations like FLUID.

**Continual Learning Analysis** We evaluate Learning without Forgetting (LwF) and Elastic Weight Consolidation (EWC). For analysis, we compare to the baselines NCM, standard training, fine-tuning. We evaluate LwF and EWC because they are well-known CL methods that are general enough to be reasonably adapted to the FLUID setup. We find that fine-tuning the classifier outperforms CL methods in both accuracy and mean per-class accuracy in FLUID.

Note that FLUID can be considered a specific instance of the CL formulation with the significant difference from standard CL formulations being the inclusion of pretraining, flexible training, and not batching as discussed in 2. We contend pretraining is a reasonable inclusion as real-world vision systems have access to large datasets such as ImageNet. From the experiments, we find that initializing networks with standard pretraining almost completely mitigates the effect of catastrophic forgetting in its current formulation, indicating we may need to rethink current evaluations.

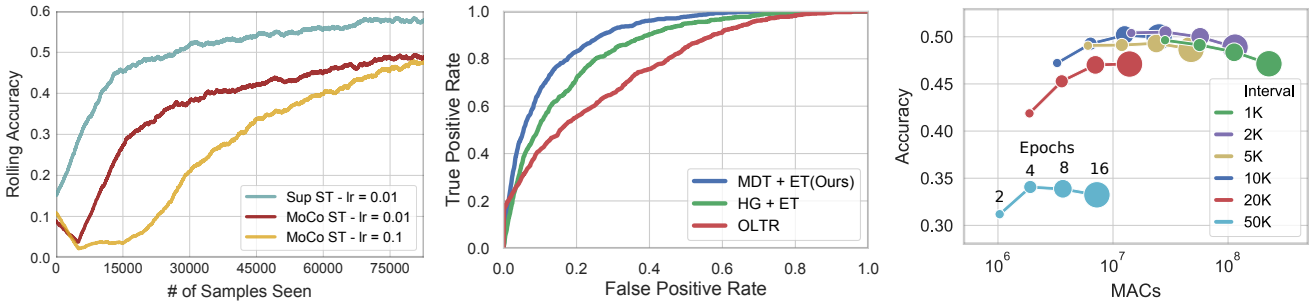
We find that NCM and fine-tuning outperform LwF and EWC on both pretrain and new classes which indicates that continual learning methods are limited in situations with large-scale pretraining. NCM uses no memory caching or replay buffer and has frozen features while LwF and EWC cache all stream images and freely train their features. NCM learns new classes as well as the CL methods with no possibility of forgetting. Furthermore, fine-tuning which also uses frozen features learned during pretraining outperforms both methods on novel and old classes and has $\sim 4\%$ higher overall accuracy. This finding validates the argument that the details of benchmarks directly impact the design and utility of the methods developed for them; thus making a case for FLUID like frameworks that unify multiple setups and provide a more holistic evaluation. We also acknowledge that for scenarios in which the pretraining data is radically different from the target distribution the above conclusion may not hold, such as for permuted MNIST.

**Exemplar Tuning** We find that ET (Table 3-l) has significantly higher overall and mean-class accuracy than all other evaluated methods and uses similar compute as fine-tuning. Figure 2-a shows how ET quickly adapts to new classes and continues to learn in the standard data regime (high accuracy at the start and end of the stream). Finally, we show that ET outperforms simple NCM + fine-tuning (Weight Imprinting) by $\sim$10%, in addition to the practical advantages outlined in section 4.

ET is a simple, general baseline that works not only for FLUID, as evidenced through its effectiveness in standard recognition tasks on Mini-ImageNet [58] & ImageNet-LT [36] (see Appendix M).

**New Class Detection and MDT** We measure AUROC for detecting new classes throughout the sequence and present in Figure 3-b. HG baseline + ET, OLTR, and MDT + ET achieve 0.84, 0.78 and 0.92 AUROC scores respectively. The performance of Minimum Distance Thresholding (MDT) indicates that standard recognition networks are well suited for detecting out-of-distribution classes and can be done simultaneously with classification. We compare MDT and HG baseline with other classifiers such as NCM and fine-tuning in Appendix K.

**Representation Learning** We observe unexpected behavior from contrastive MoCo [20] and VINCE [15] representations in the FLUID setting . For example, when fine-tuning the classifier of a MoCo representation we find that accuracy is less than 1% and 4.96% on novel and pretrain tail classes respectively (Table 3-o). In comparison the supervised counterpart (Table 3-i) obtains 23.56% and 58.77% accuracy respectively. We conjecture that this difficulty is induced by learning with a linear classifier as NCM with MoCo pretraining (Table 3-q) does not exhibit the same difficulties as standard training and fine-tuning. Figure 3-a shows




(a) (b) (c)

Figure 3. (a) Accuracy of standard training with MoCo & supervised pretraining. Surprisingly, MoCo accuracy falls during initial streaming phase. (b) ROC curves for unseen class detection. MDT outperforms all OOD baselines evaluated in FLUID. (c) Standard training accuracy curve for a range of training frequencies & epochs showing that over training can lead to lower accuracy. MACs $\propto$ total gradient updates.

other unexpected behavior where MoCo accuracy decreases to almost 0% initially when standard training, then begins improving after 10K samples. This behavior is not observed for supervised pretraining and occurs for multiple learning rates. We argue that this is related to learning a mixture of pretrain and novel classes which is the primary difference between FLUID and previous downstream tasks. The significantly lower accuracy of MoCo representations on novel tail classes while fine-tuning (Table 3-o) further reinforces this hypothesis. These observations and insights are further supported by similar results obtained for representations trained via VINCE [15], a self-supervised contrastive representation learning method (see Appendix G). These results validate the utility of a unified evaluation such as FLUID which assess the capabilities of methods along more dimensions.

**Update Strategies** We briefly investigate trade-offs between compute cost and accuracy of simple update strategies. The challenge of learning adaptive update strategies is an unexplored problem posed by FLUID left to future work.

We evaluate the accuracy and total compute cost of combinations of varying update frequencies and number of training epochs per update (Figure 3-c & 6). We conduct our experiments for fine-tuning (Figure 6 in Appendix N) and for standard training (Figure 3-c) on ResNet18 model with supervised pretraining.

We observe that training for too many total epochs (training frequency $\times$ epochs) with standard training (Figure 3-c) unexpectedly decreases overall accuracy, though fine-tuning asymptotically improves (Figure 6 in Appendix N). We hypothesize that the optimal amount of training balances the features learnt from ImageNet-1K with those from the smaller, imbalanced streaming data. This aligns with our continual learning experiments that indicate large-scale pretrained features trained on more data outperform specialized features. These initial experiments are intended to illustrate the new problems that FLUID presents for future research. The results indicate that there is significant room for improvement in both efficiency and accuracy with new strategies for training networks under streaming conditions which we leave for future work.

## 6. Limitations and Future Work

Throughout this paper, we studied various methods and settings in the context of supervised image classification, a highly explored problem in ML. While we do not make design decisions specific to image classification, incorporating other mainstream tasks into FLUID is an immediate next step. Also while the FLUID framework is agnostic to any particular data set, our experiments and conclusions are anchored in the ImageNet domain. Across the experiments in this paper, we impose some assumptions about the learning conditions, albeit only a few, on FLUID. For example, we currently assume that FLUID has access to labels as the data streams in. One exciting future direction is to add the semi- and un-supervised aspects to FLUID. Relaxing these remaining assumptions to bring FLUID even closer to real world settings is an interesting direction for future work.

## 7. Conclusions

We introduce FLUID, a unified evaluation framework designed to facilitate research towards more general methods capable of handling the challenges of learning in practical settings. FLUID enables comparison and integration of solutions across few-shot, transfer, continual representation learning, & out-of-distribution detection while introducing new research challenges like how and when to update model parameters based on incoming data. Through our experiments with FLUID on a wide-range of methods we show the limitations and merits of existing solutions. For example, canonical few-shot methods do not scale well to more classes/examples and continual learning techniques reduce accuracy with large-scale pretraining. As a starting point for developing general methods on FLUID, we present two simple baselines, Exemplar Tuning & Minimum Distance Thresholding, which outperform all other evaluated methods.

## Acknowledgements

This work is in part supported by NSF IIS 1652052, IIS 17303166, DARPA N66001-19-2-4031, 67102239 and gifts from Allen Institute for Artificial Intelligence. We thank Jae Sung Park and Mitchell Wortsman for insightful discussions and Daniel Gordon for the pretrained MoCo weights.

## References

[1] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019. 2

[2] Rahaf Aljundi, Marcus Rohrbach, and Tinne Tuytelaars. Selfless sequential learning. *arXiv preprint arXiv:1806.05421*, 2018. 2

[3] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015. 4

[4] Albert Bifet, Geoff Holmes, Richard Kirkby, and Bernhard Pfahringer. Moa: Massive online analysis. *Journal of Machine Learning Research*, 11:1601–1604, 2010. 3

[5] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 5

[6] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pages 52–68. Springer, 2016. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 4

[8] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020. 5

[9] Jaeik Cho, Taeshik Shon, Ken Choi, and Jongsub Moon. Dynamic learning model update of hybrid-classifiers for intrusion detection. *The Journal of Supercomputing*, 64(2):522–526, 2013. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 12

[11] Vincent Dumoulin, Neil Houlsby, Utku Evci, Xiaohua Zhai, Ross Goroshin, Sylvain Gelly, and Hugo Larochelle. Comparing transfer and meta learning approaches on a unified few-shot classification benchmark. *arXiv preprint arXiv:2104.02638*, 2021. 3

[12] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. Rigging the lottery: Making all tickets winners. In *Proceedings of the International Conference on Machine Learning*, 2020. 3

[13] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017. 5, 6

[14] Heitor Murilo Gomes, Jesse Read, Albert Bifet, Jean Paul Barddal, and João Gama. Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explorations Newsletter*, 2019. 3

[15] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 4, 7, 8, 13, 14, 15

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 4

[17] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3018–3027, 2017. 3

[18] James Harrison, Apoorva Sharma, Chelsea Finn, and Marco Pavone. Continuous meta-learning without tasks. *Advances in neural information processing systems*, 2019. 2

[19] Jiangpeng He, Runyu Mao, Zeman Shao, and Fengqing Zhu. Incremental learning in online scenario. *arXiv preprint arXiv:2003.13191*, 2020. 2

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019. 4, 6, 7, 13, 14

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 4, 5

[23] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[24] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

[25] 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*, 2017. 2, 5, 13

[26] Shu Kong and Deva Ramanan. Opengan: Open-set recognition via open data generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 4

[27] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 4

[28] Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. Soft threshold weight reparameterization for learnable sparsity. In *Proceedings of the International Conference on Machine Learning*, 2020. 3

[29] Aditya Kusupati, Matthew Wallingford, Vivek Ramanujan, Raghav Somani, Jae Sung Park, Krishna Pillutla, Prateek Jain, Sham Kakade, and Ali Farhadi. Llc: Accurate, multi-purpose learnt low-dimensional binary codes. In *Advances in neural information processing systems*, 2021. 3

[30] Brenden Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua Tenenbaum. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33, 2011. 4

[31] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 4

[32] Kimin Lee. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in neural information processing systems workshop*. 4, 5

[33] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 2, 5, 13

[34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[35] Zhiqiu Lin, Deva Ramanan, and Aayush Bansal. Streaming self-training via domain-agnostic unlabeled images. *arXiv preprint arXiv:2104.03309*, 2021. 3

[36] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 4, 5, 7, 13, 14

[37] Marc Masana, Idoia Ruiz, Joan Serrat, Joost van de Weijer, and Antonio M Lopez. Metric learning for novelty and anomaly detection. In *Proceedings of the British Machine Vision Conference*, 2018. 4, 5

[38] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018. 3, 5, 6

[39] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019. 13

[40] Ameya Prabhu, Philip H.S. Torr, and Puneet K. Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *Proceedings of the European Conference on Computer Vision*, 2020. 3

[41] Senthil Purushwalkam and Abhinav Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020. 4

[42] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 5, 14

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 4

[44] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*, pages 525–542. Springer, 2016. 3

[45] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. *International Conference on Learning Representations*, 2017. 3

[46] 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*, pages 2001–2010, 2017. 2

[47] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing inference. *Proceedings of*

*the IEEE conference on computer vision and pattern recognition*, 2019. 2

[48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3, 4, 12

[49] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 5, 6, 14, 16

[50] Jerzy Stefanowski and Dariusz Brzezinski. Stream classification., 2017. 3

[51] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–412, 2019. 3, 5, 6

[52] Yu Sun, Xiaolong Wang, Liu Zhuang, John Miller, Moritz Hardt, and Alexei A. Efros. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 2

[53] Sebastian Thrun. Is learning the n-th thing any easier than learning the first? *Advances in neural information processing systems*, 1996. 2

[54] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020. 5

[55] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018. 4

[56] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 4

[57] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021. 4

[58] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 4, 7, 14

[59] Yan Wang, Wei-Lun Chao, Kilian Q Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *arXiv preprint arXiv:1911.04623*, 2019. 5

[60] Kapil K Wankhade, Snehlata S Dongre, and Kalpana C Jondhale. Data stream classification: a review. 3

[61] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020. 4

[62] Davis Wertheimer and Bharath Hariharan. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6558–6567, 2019. 4

[63] Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. Supermasks in superposition. In *Advances in Neural Information Processing Systems*, 2020. 2

## A. FLUID Procedure

Algorithm 1 describes the high level implementation directives of FLUID framework.

**Algorithm 1** FLUID Procedure

**Input:** Task $\mathcal{T}$
**Input:** ML sys.: (pretrained) model $\mathbf{f}$, update strategy $\mathbf{S}$
**Output:** Evaluations: $E$, Operation Counter: $C$

**function** FLUID($\mathcal{T}$, ($\mathbf{f}$, $\mathbf{S}$))
  Evaluations $E = [\,]$
  Datapoints $\mathcal{D} = [\,]$
  Operation Counter $C = 0$.

  **while** streaming **do**
    Sample $\{x, y\}$ from $\mathcal{T}$
    prediction $p = \mathbf{f}(\mathbf{x})$ ($A$ operations)
    Flag $n$ indicates if $y$ is a new unseen class
    $E$.insert($\{y, p, n\}$)
    $\mathcal{D}$.insert($\{x, y\}$)
    Update $\mathbf{f}$ using $\mathbf{S}$ with $\mathcal{D}$ ($B$ operations)
    $C$ += $A + B$
  **end while**

  **return** $E, C$
**end function**

## B. Dataset Information

The five sequences we pair with FLUID are constructed from ImageNet-22K [10]. Two sequences (1-2) are for validation, and three (3-5) are for testing. Each sequence contains 1,000 classes; 250 of which are in ImageNet-1K [48] (pretrain classes) and 750 of which are only in ImageNet-22K (novel classes). For the test sequences, we randomly select the classes without replacement to ensure that the sequences do not overlap. The validation sequences share pretrain classes because there are not enough pretrain classes (1000) to partition among five sequences. We randomly distribute the number of images per class according to Zipf's law with $s = 1$ (Figure 4). For classes without enough images, we fit the Zipfian distribution as closely as possible which causes a slight variation in sequence statistics seen in Table 4.

Table 4. Statistics for the sequences of images used in FLUID. Sequences 1-2 are for validation and Sequence 3-5 are for testing. The images from ImageNet-22k are approximately fit to a Zipfian distribution with 250 classes overlapping with ImageNet-1k and 750 new classes.

| Sequence # | Number of Images | Min # of Class Images | Max # of Class Images |
|---|---|---|---|
| 1 | 89030 | 1 | 961 |
| 2 | 87549 | 21 | 961 |
| 3 | 90133 | 14 | 961 |
| 4 | 86988 | 6 | 892 |
| 5 | 89921 | 10 | 961 |

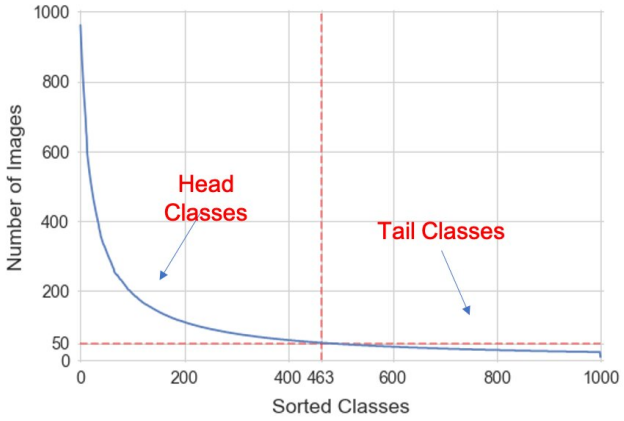



Figure 4. The distribution of samples over the classes for Sequences 1 - 5. Classes with less than 50 samples are considered in the tail and samples with greater than or equal to 50 samples are considered in the head for the purpose of reporting.

## C. Dataset License

ImageNet does not explicitly provide a license.

## D. More Method Details

**Nearest Class Mean** Each class mean, $m_i$, is the average feature embedding of all examples in class $i$: $m_i = \sum_{x \in C_i} f_\phi(x)$; where $C_i$ is the set of examples belong to class $i$ and $f_\phi$ is the deep feature embedding of $x$. Class probabilities are the softmax of negative distances between $x$ and class means:

$$P(y = i|\mathbf{x}) = \frac{e^{-d(m_i, f_\phi(\mathbf{x}))}}{\sum_{i'} e^{-d(m_{i'}, f_\phi(\mathbf{x}))}} \tag{2}$$

**MAML** The gradient update for MAML is: $\theta \leftarrow \theta - \beta \cdot \nabla_\theta \sum_{\mathcal{T}_i \sim p(T)} \mathcal{L}_{\mathcal{T}_i}\left(f_{\theta_i'}\right)$ where $\theta_i'$ are the parameters after making a gradient update given by: $\theta_i' \leftarrow \theta - \alpha \cdot \nabla_\theta \mathcal{L}_{\mathcal{T}_i}\left(f_\theta\right)$.

**OLTR** The network consist of two parts 1) A feature extractor consist of a ResNet backbone followed by a modulated attention and 2) A classifier and memory bank that are used to classify the output of the feature extractor. Training is done in 2 stages; In the first stage the feature extractor is trained. In the second stage the feature extractor and classifier are fine-tuned while samples are accumulated in the memory bank.

**Weight Imprinting** Weight Imprinting initializes the weights of the cosine classification layer, then performs fine-tuning using all of the data with a learnable temperature to scale the logits. Weight Imprinting can be thought of as NCM with cosine similarity as the metric for determining the closest neighbor, then performing fine-tuning. To use Weight Imprinting in a sequential setting, rather than a few-shot setting, we must decide when to begin fine-tuning. In

the original formulation, fine-tuning was performed after the centroids were calculated using the entire data set, but in the sequential setting we do not have access to the entire data set until streaming ends. Therefore we choose to begin fine-tuning when the accuracy of fine-tuning exceeds NCM on validation data. In a real-world scenario it would be difficult to obtain such information, but we employ such a strategy to provide an upper-bound for the performance of Weight Imprinting in the sequential setting.

## E. Implementation Details

In this section, we discuss how methods are adapted with respect to FLUID. Some methods are intuitively applied with little modification, and some require interpretation for how they should be adapted.

**Offline Training** For all experiments (Table 3) that require offline training (fine-tuning, Weight Imprinting, standard training, ET and LwF), except OLTR, we train each model for 4 epochs every 5,000 samples observed. An epoch includes training over all previously seen data in the sequence. Experiments in Figure 6 show that training for 4 epochs every 5,000 samples balanced sufficient accuracy and reasonable computational cost. Fine-tuning experiments use a learning rate of 0.1 and standard training uses 0.01 for supervised pretraining. For MoCo pretraining fine-tuning uses a learning rate of 30 and standard training uses 0.01. All the experiments use the SGD+Momentum optimizer with a 0.9 momentum.

**Instance-Based Updates** All instance-based methods (NCM, ET, Weight Imprinting, Prototypical Networks) are updated after every sample as it takes no additional compute compared to batch updates.

**Meta-Training** For few-shot methods that leverage meta-training for pretraining, we used 5-shot 30-way except for MAML which we meta-trained with 5-shot 5-way due to computational costs. We choose to use 30-way as the computational graph is limited in the number of instances that it can store in memory and backpropagate to. We meta-train for 100 epochs with a learning rate of 0.01 and reduce it by 0.5 every 40 epochs.

**Exemplar Tuning** We initialize the residual vectors as zero. ET is trained according to the specifications of instance-based updates and offline training simultaneously.

**Weight Imprinting** For Weight Imprinting, we transition from NCM to fine-tuning after 10,000 samples as we observed that the accuracy of NCM saturated at this point in the validation sequence. We use a learning rate of 0.1 while fine-tuning.

**Learning Without Forgetting** We adapt Learning Without Forgetting to the FLUID task by freezing a copy of the model after pretraining which is used for knowledge distillation. Not all pretraining classes are seen during streaming so only softmax probabilities for classes seen during the stream are used in the cross-entropy between the soft labels and predictions. We use a temperature of 2 to smooth the probabilities in accordance with [33]. Training is done according to the specifications given in the offline training portion of this section.

**Elastic Weight Consolidation** We adapt Elastic Weight Consolidation [25] to the FLUID task by freezing a copy of the model after pretraining as the optimla model of the pretrain task. We then use the validation set of ImageNet-1K corresponding to the 250 classes being used for the computation of Fisher information per-parameter. For every training step in FLUID, a penalty is added based on the distance moved by the parameters from the base model weighted by the Fisher information. The Fisher information is calculated at the start of every flexible train step to mitigate catastrophic forgetting. Training is done according to the specifications given in the offline training portion of this section. Depending on how frequently the Fisher information is computed, the compute associated increases over the standard training costs.

**OLTR** For OLTR [36], we update the memory and train the model for 4 epochs every 200 samples for the first 10,000 samples, then train 4 epochs every 5,000 samples with a 0.1 learning rate for classifier parameters and 0.01 for feature extraction parameters which is in accordance with the specifications of the original work.

**Pretraining** We use the PyTorch [39] ResNet18 and ResNet50 models pretrained on supervised ImageNet-1K. We use the models from Gordon et al. [15] for the MoCo [20] self-supervised ImageNet-1K pretrained models. MoCo-ResNet18 and MoCo-ResNet50 get top-1 validation accuracy of 44.7% and 65.2% respectively and were trained for 200 epochs. For fine-tuning and ET with MoCo, we report the results with a learning rate of 30 which is suggested by the original work when learning on frozen features. All other learning rates with MoCo are the same as with supervised.

## F. Training Depth for Fine Tuning

We explored how training depth affects the accuracy of a model on new, old, common, and rare classes. For this set of experiments, we vary the number of trained layers when fine-tuning for 4 epochs every 5,000 samples on ResNet18 with a learning rate of 0.01 on Sequence 2 (validation). The results are reported in Table 5. We found that training more layers leads to greater accuracy on new classes and lower accuracy on pretrain classes. However, we observed that the number of fine-tuning layers did not significantly affect overall accuracy so for our results on the test sequences (3-5) we only report fine-tuning of one layer (Table 3).

Table 5. The results for fine-tuning various numbers of layers with a learning rate of .01 on Sequence 2. Training more layers generally results in higher accuracy on novel classes, but lower accuracy on pretrain classes. The trade-off between novel and pretrain accuracy balances out so the overall accuracy is largely unaffected by the depth of training.

| # of Layers | Novel-Head (>50) | Pretrain-Head (>50) | Novel-Tail (<50) | Pretrain-Tail (<50) | Mean Per-Class | Overall |
|---|---|---|---|---|---|---|
| 1 | **41.32** | **80.96** | 17.13 | 66.52 | 39.19 | 56.87 |
| 2 | 41.55 | 80.79 | 17.40 | **67.03** | 39.43 | 56.79 |
| 3 | 45.82 | 78.59 | 19.08 | 59.52 | **40.73** | **57.23** |
| 4 | **46.96** | 75.44 | 19.87 | 53.97 | 40.39 | 57.04 |
| 5 | 46.76 | 75.72 | **19.97** | 54.04 | 40.41 | 57.04 |

## G. Results for VINCE ResNet18 backbone on Sequence 5

We report all performance metrics for sequence 5 in Table 6 for ResNet18 backbone trained via VINCE [15]. VINCE is a self-supervised representation learning method that focuses on leveraging video as a natural form of augmentation for contrastive learning. These results corroborate the findings of Table 3 which uses ResNet18 backbone trained via MoCo [20] further solidifying the insights drawn on self-supervised representation learning methods.

## H. Results for ResNet50 backbone on Sequence 5

We report all performance metrics for sequence 5 in Table 7 for ResNet50 backbone. These results corroborate the findings of Table 3 which uses ResNet18 backbone.

## I. Results For Other Sequences

We report the mean and standard deviation for all performance metrics across test sequences 3-5 in Table 8. Note that the standard deviation is relatively low so the methods are consistent across the randomized sequences.

## J. Prototypical Network Experiments

We benchmarked our implementation of Prototypical Networks on few-shot baselines to verify that it is correct. We ran experiments for training on both MiniImageNet [58] and regular ImageNet-1k and tested our implementation on the MiniImageNet test set and FLUID (Sequence 2). We found comparable results to those reported by the original Prototypical Networks paper [49] (Table 9).

## K. Out-of-Distribution Ablation

In this section we report AUROC and F1 for MDT and softmax for all baselines. In section 5 we only included OLTR, MDT with Exemplar Tuning, and ET with maximum softmax (Hendrycks Baseline). Additionally, we visualize the accuracy curves for in-distribution and out-of-distribution samples as the rejection threshold vary (Figure 5). All the OOD experiments presented in Figure 5 and Table 10 were run using ResNet18. Minimum Distance Thresholding (MDT) threshold distances but also similarity metrics can be used. MDT generally works better than maximum softmax when applied to most methods.

The results of NCM and Exemplar Tuning using softmax and dot product similarity in comparison to OLTR are shown in table 10. The F1-scores are low due to the large imbalance between positive and negative classes. There are 750 unseen class datapoints vs $\sim$ 90000 negative datapoints. Table 10 shows that cosine similarity (MDT) is better than softmax or the OLTR model for most methods.

## L. Weight Imprinting and Exemplar Tuning Ablations

In Table 11, we ablate over various softmax temperature initializations with Weight Imprinting. We learn the temperature as described in [42], but find that initial value affects performance. We report the best results in the main paper. We also ablate over the similarity metrics use in ET. We find that the dot product (linear) is the best measure of similarity for ET.

## M. Exemplar Tuning on Standard Recognition Tasks

On Mini-ImageNet [58], for 5-shot 5-way Exemplar Tuning with a ResNet10 backbone obtains an accuracy 72.1% compared to 68.2% for Prototypical Networks. Exemplar Tuning accuracy on ImageNet-LT [36] with a ResNet18 backbone is 42.1% while a standard linear layer gets to 41.9%.

## N. Update Strategies

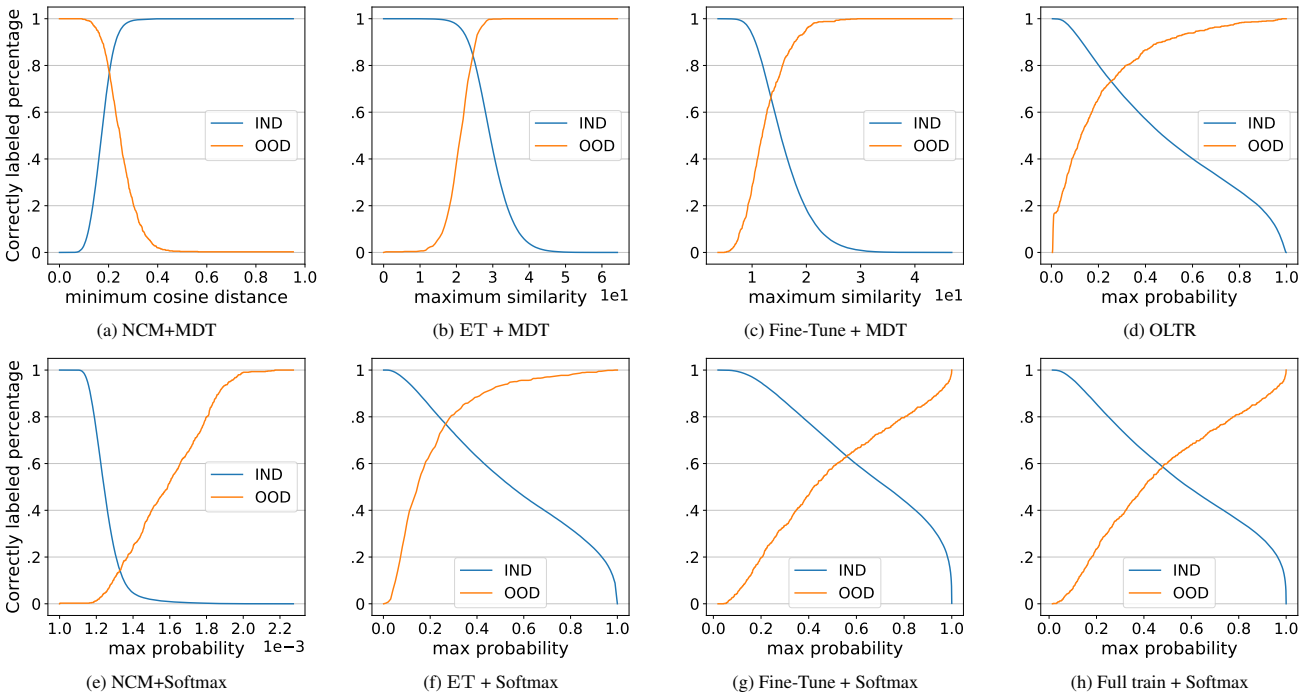
Figure 6 has the accuracy vs MACs trade-off for fine-tuning across various update strategies.

Table 6. Results on sequence 5 with ResNet18 backbone trained using VINCE [15]

| Method | Pretrain Strategy | Novel - Head (>50) | Pretrain - Head (>50) | Novel - Tail (<50) | Pretrain - Tail (<50) | **Mean Per-Class** | **Overall** | GMACs↓ ($\times 10^6$) |
|---|---|---|---|---|---|---|---|---|
| Backbone - ResNet18 | | | | | | | | |
| (a) Fine-tune | VINCE | 18.00 | 14.61 | 1.89 | 1.56 | 7.25 | 26.27 | 0.16 / 5.73 |
| (b) Standard Training | VINCE | 24.60 | 32.06 | 6.38 | 9.63 | 16.17 | 32.95 | 11.29 |
| (c) NCM | VINCE | 15.96 | 22.32 | **12.32** | **16.08** | 15.20 | 18.28 | **0.15** |
| (d) **Exemplar Tuning** | VINCE | **26.84** | **37.03** | 7.32 | 11.30 | **18.11** | **35.44** | 0.16 / 5.73 |

Table 7. Continuation of Table 3 results on sequence 5 with ResNet50 backbone.

| Method | Pretrain Strategy | Novel - Head (>50) | Pretrain - Head (>50) | Novel - Tail (<50) | Pretrain - Tail (<50) | **Mean Per-Class** | **Overall** | GMACs↓ ($\times 10^6$) |
|---|---|---|---|---|---|---|---|---|
| Backbone - ResNet50 | | | | | | | | |
| (a) Fine-tune | MoCo | 14.42 | 43.61 | 0.22 | 13.40 | 11.85 | 31.35 | 0.36 / 13.03 |
| (b) Fine-tune | Sup. | 47.78 | 82.06 | 27.53 | **66.42** | 46.24 | 57.95 | 0.36 / 13.03 |
| (c) Standard Training | MoCo | 26.82 | 42.12 | 10.50 | 21.08 | 21.32 | 35.44 | 38.36 |
| (d) Standard Training | Sup. | 43.89 | 74.50 | 21.54 | 50.69 | 39.48 | 54.10 | 38.36 |
| (e) NCM | MoCo | 30.58 | 55.01 | 24.10 | 45.37 | 32.75 | 36.14 | **0.35** |
| (f) NCM | Sup. | 45.58 | 78.01 | **35.94** | 62.90 | 47.75 | 52.19 | **0.35** |
| (g) LwF | Sup. | 21.52 | 49.17 | 5.49 | 38.74 | 20.69 | 30.57 | 38.36/76.72 |
| (h) EWC | Sup. | 43.84 | 76.03 | 21.22 | 53.64 | 39.89 | 54.59 | 40.36 |
| (i) **Exemplar Tuning** | MoCo | 28.86 | 54.03 | 7.02 | 20.82 | 21.89 | 40.13 | 0.36 / 13.03 |
| (j) **Exemplar Tuning** | Sup. | **52.95** | **82.27** | 28.13 | 57.15 | **48.02** | **62.41** | 0.36 / 13.03 |




(a) NCM+MDT (b) ET + MDT (c) Fine-Tune + MDT (d) OLTR

(e) NCM+Softmax (f) ET + Softmax (g) Fine-Tune + Softmax (h) Full train + Softmax

Figure 5. The accuracy for the in-distribution (IND) and out-of-distribution (OOD) samples as the threshold for considering a sample out-of-distribution varies. The horizontal axis is the threshold value, and the vertical axis is the accuracy. Intersection of the IND and OOD curves at a higher accuracy generally indicates better out-of-distribution detection for a given method.

Table 8. Averaged results for all methods evaluated on Sequences 3-5. See Table 3 for the computational cost (GMACs) for each method and more information about each column.

| Method | Pretrain | Novel - Head (>50) | Pretrain - Head (>50) | Novel - Tail (<50) | Pretrain - Tail (>50) | **Mean Per-Class** | **Overall** |
|---|---|---|---|---|---|---|---|
| Backbone - Conv-4 | | | | | | | |
| Prototype Networks | Meta | 5.02±0.05 | 9.71±0.11 | 0.64±0.01 | 1.27±0.04 | 3.25±0.03 | 7.82±0.09 |
| MAML | Meta | 2.93±0.01 | 2.02±0.02 | 0.15±0.01 | 0.1±0.01 | 1.11±0.02 | 3.64±0.06 |
| Backbone - ResNet18 | | | | | | | |
| Prototype Networks | Meta | 8.72±0.09 | 16.84±0.14 | 7.06±0.03 | 12.98±0.04 | 9.46±0.08 | 11.19±0.12 |
| Meta-Baseline | Sup./Meta | 41.73±0.57 | 66.54±2.37 | 27.54±1.13 | 53.69±0.97 | 39.32±0.71 | 47.74±0.63 |
| Fine-tune | Moco | 5.31±0.24 | 45.95±1.27 | 0.03±0 | 26.23±0.88 | 10.64±0.23 | 18.52±0.98 |
| Fine-tune | Sup. | 43.2±0.65 | 74.55±2.53 | 22.79±1.21 | 59.63±1.02 | 40.9±0.73 | 53.06±0.65 |
| Standard Training | Moco | 26.9±0.27 | 42.39±3.04 | 9.1±0.74 | 21.11±0.51 | 20.76±0.32 | 34.85±0.75 |
| Standard Training | Sup. | 38.82±0.49 | 65.88±2.32 | 16.15±0.83 | 44.3±0.91 | 33.63±0.38 | 48.81±0.57 |
| NCM | Moco | 19.31±0.06 | 30.02±1.69 | 14.21±0.46 | 22.06±0.52 | 18.86±0.13 | 22.14±1.24 |
| NCM | Sup. | 41.68±0.65 | 70.05±2.29 | 31.24±0.86 | 57.23±0.97 | 42.87±0.62 | 47.89±0.76 |
| OLTR | MoCo | 41.47±0.03 | 31.48±0.01 | 17.48±0.01 | 9.81±0.01 | 22.03±0 | 38.33±0.01 |
| OLTR | Sup. | 51.19±0.37 | 37.02±0.51 | 24.14±0.14 | 13.77±0.24 | 27.6±0.28 | 44.46±0.44 |
| **Exemplar Tuning** | Moco | 32.57±1.54 | 43.48±0.4 | 6.39±0.49 | 12.81±0.12 | 18.46±0.35 | 39.25±1.20 |
| **Exemplar Tuning** | Sup. | 46.36±2.31 | 69.34±0.53 | 23.48±1.23 | 45.82±0.32 | 42.93±0.17 | 57.56±0.56 |
| Backbone - ResNet50 | | | | | | | |
| Fine-tune | Moco | 45.95±0.26 | 5.31±0.32 | 26.23±0.07 | 0.03±1.74 | 10.64±0.21 | 18.52±1.02 |
| Fine-tune | Sup. | 47.59±0.65 | 80.14±1.71 | 26.69±0.97 | 66.92±1.4 | 45.62±0.6 | 57.48±0.47 |
| Standard Training | Moco | 43.93±0.73 | 71.72±3.18 | 20.84±0.92 | 51.43±0.68 | 38.94±0.9 | 53.45±1.73 |
| Standard Training | Sup. | 47.59±0.45 | 80.14±2.59 | 26.69±0.79 | 66.92±1.91 | 45.62±0.47 | 57.48±0.56 |
| NCM | Moco | 30.15±0.48 | 53.84±1.05 | 23.99±0.53 | 44.11±1.11 | 32.27±0.92 | 35.45±0.61 |
| NCM | Sup. | 45.46±0.95 | 76.55±1.77 | 35.47±0.82 | 65.62±1.57 | 47.77±0.65 | 52.22±0.55 |
| **Exemplar Tuning** | Moco | 28.46±3.04 | 40.42±1.33 | 7.57±2.15 | 14.36±4.14 | 19.54±2.63 | 32.07±2.37 |
| **Exemplar Tuning** | Sup. | 49.24±1.55 | 75.78±1.84 | 26.67±2.17 | 55.63±2.31 | 44.15±1.44 | 62.35±1.02 |

Table 9. Our implementation of Prototypical Networks on MiniImageNet & FLUID. $^{\diamond}$ Results from Snell et al. [49].

| Method | Backbone | Train Set | MiniImageNet 5 Way - 5 Shot | FLUID |
|---|---|---|---|---|
| Prototypical Networks | Conv - 4 | MiniImageNet | 69.2 | 14.36 |
| Prototypical Networks | Conv - 4 | ImageNet (Train) | 42.7 | 15.98 |
| Prototypical Networks$^{\diamond}$ | Conv - 4 | MiniImageNet | 68.2 | - |

Table 10. The out-of-distribution performance for each method on sequence 5. We report the AUROC and the F1 score achieved by choosing the best possible threshold value.

| Metric | NCM +Softmax | NCM +MDT | Exemplar Tuning +Softmax | Exemplar Tuning +MDT | Standard Training +Softmax | Standard Training +MDT | Fine-Tune +Softmax | Fine-Tune +MDT | OLTR |
|---|---|---|---|---|---|---|---|---|---|
| AUROC | 0.07 | 0.85 | 0.84 | 0.92 | 0.59 | 0.53 | 0.68 | 0.72 | 0.78 |
| F1 | 0.01 | 0.20 | 0.10 | 0.20 | 0.03 | 0.02 | 0.06 | 0.10 | 0.27 |

Table 11. Comparison of Weight Imprinting and Exemplar Tuning with different classifiers and initial temperatures. Exemplar Tuning with a linear layer performs significantly better than all other variants.

| Method | Pretrain | Backbone | Novel - Head (>50) | Pretrain - Head (>50) | Novel - Tail (<50) | Pretrain - Tail (>50) | Mean Per-Class | Overall |
|---|---|---|---|---|---|---|---|---|
| Weight Imprinting (s = 1) | Sup | R18 | 36.58 | 63.39 | 9.32 | 21.80 | 26.85 | 46.35 |
| Weight Imprinting (s = 2) | Sup | R18 | 36.58 | 63.39 | 9.32 | 21.80 | 26.85 | 46.35 |
| Weight Imprinting (s = 4) | Sup | R18 | 40.32 | 67.46 | 15.35 | 34.18 | 32.69 | 48.51 |
| Weight Imprinting (s = 8) | Sup | R18 | 31.18 | 32.66 | 34.77 | 28.94 | 32.56 | 46.67 |
| Exemplar Tuning (Cosine) | Sup | R18 | 33.90 | 18.22 | 4.84 | 1.88 | 11.72 | 31.81 |
| Exemplar Tuning (Euclidean) | Sup | R18 | 43.40 | 66.32 | 21.66 | 42.06 | 37.19 | 51.62 |
| Exemplar Tuning (Linear) | Sup | R18 | **48.85** | **75.70** | **23.93** | **45.73** | **43.61** | **58.16** |

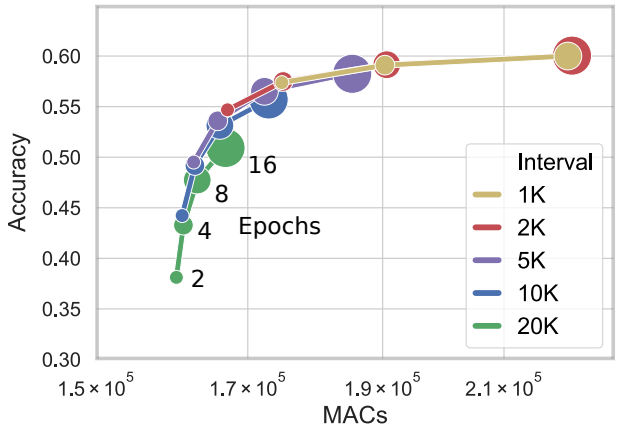



Figure 6. The plot compares the accuracy and MACs for various update strategies when fine-tuning.