WHEN THE TEACHER LAGS: CONTRASTIVE PRE-TRAINING UNDER DISTRIBUTION DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive pre-training typically assumes that training data follow a stable distribution. When the distribution changes, however, a momentum teacher may retain information from earlier data and provide targets that are poorly aligned with the current samples. We study how this lag affects representation learning and use a structural causal model to characterize the resulting source of bias. Guided by this analysis, we propose Resilient Contrastive Pre-training (RCP), which introduces an intervention-inspired contrastive objective to reduce the influence of drift-induced bias during training. RCP integrates a query–key–value module into a momentum-based student–teacher framework and learns from samples within a drift adaptation window. We evaluate the resulting representations through long-tailed classification, domain-shifted classification, and out-of-distribution detection. Across these evaluations, RCP improves upon standard contrastive pre-training baselines, with particularly consistent gains for underrepresented classes. These findings suggest that accounting for drift during pre-training can improve the robustness and transferability of learned visual representations. Codes are available at `https://anonymous.4open.science/r/ResilientCL/`.

1 INTRODUCTION

Contrastive learning has emerged as a dominant paradigm for large-scale visual pre-training, demonstrating impressive performance across a wide range of vision tasks (He et al., 2020; Caron et al., 2021; Zhou et al., 2024). While most existing studies evaluate contrastive methods on static, fully-scraped internet corpora, the pre-training landscape is shifting toward dynamic and continuously expanding data regimes(Yang et al., 2024a; 2025a). Modern large-scale systems no longer rely on a one-shot frozen dataset. Instead, they accumulate data over time from heterogeneous sources such as user-generated content in federated/edge-learning pipelines(Gao et al., 2022; Chen et al., 2023; Panchal et al., 2023), medical imaging networks (Jiang et al., 2022; Chen et al., 2024; Yang et al., 2022; 2024b; 2021a), and autonomous-driving fleets (Liang et al., 2022; Zheng et al., 2023). In these real-world settings, data arrive sequentially, exhibit strong non-IID characteristics, and naturally undergo concept drift, including evolving class frequencies, shifting domains, seasonal changes, and sensor-induced feature heterogeneity (Lu et al., 2019; Yang et al., 2024a; 2025b). Recent studies in federated learning (Chen et al., 2023; Jiang et al., 2022; Panchal et al., 2023) and autonomous driving (Zheng et al., 2023; Liang et al., 2022) consistently highlight that drifting data distributions degrade representation quality and destabilize pre-trained backbones. These developments raise a critical but under-explored question:

Can contrastive pre-training reliably learn from non-stationary and drifting data?

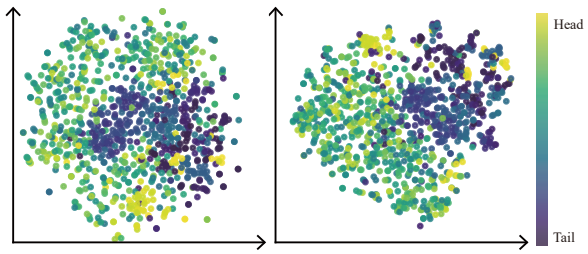
Through theoretical analysis and empirical evidence, we show that current contrastive objectives are highly sensitive to temporal distribution shift, leading to biased momentum teachers, collapsed similarity structures, and unstable representation dynamics. This exposes an urgent need for principled approaches that make contrastive pre-training robust to drift in realistic, continuously evolving environments. A review of related work is provided in the Appendix.

Current contrastive pre-training methods primarily rely on maximizing the agreement between two augmented views of the same input instance, typically processed through different encoders or network states (Chen et al., 2020; Grill et al., 2020; He et al., 2020). In the context of large vision models, this paradigm is often instantiated through student–teacher architectures, as exemplified by

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.

frameworks such as DINO (Caron et al., 2021) and MoCo v3 (Chen et al., 2021). In these setups, the student network is optimized via a contrastive objective, while the teacher network is updated using an exponential moving average (EMA) of the student parameters to stabilize training. However, this widely adopted momentum-based update mechanism becomes particularly vulnerable under concept drift. As the underlying data distribution evolves, the teacher network—updated only gradually through EMA—tends to retain outdated knowledge from previous distributions. This stale information accumulates over time, propagating outdated representations and amplifying the misalignment between the learned feature space and the current data characteristics, ultimately degrading the quality of the pre-trained representations.

To concretely illustrate this phenomenon, we examine the effect of long-tailed drift—a prevalent form of concept drift where class frequencies evolve over time—on the contrastive pre-training process. We visualize the feature space learned by MoCo v3 (Chen et al., 2021) in Figure 1. We simulate the pre-training process on two scenarios: (a) standard pre-training on the balanced ImageNet dataset (Russakovsky et al., 2015); and (b) drift pre-training on ImageNet-LT (Liu et al., 2019), which simulates a non-stationary stream with an increasingly imbalanced distribution. For a fair comparison, the feature embeddings are projected using t-SNE and evaluated on a balanced test set. Category colors range from dark to light, representing the transition from tail classes (few training samples) to head classes (abundant samples). As shown in Figure 1, the feature space under long-tailed drift becomes highly distorted: tail categories collapse into a compact region, while head categories dominate most of the embedding manifold. This imbalance contrasts sharply with the well-structured and evenly distributed feature space obtained under balanced pre-training. Such degradation can be attributed to the bias accumulation in the momentum-updated teacher network, which is repeatedly exposed to head-class samples. As the distribution drifts, the teacher provides outdated or skewed supervision, leading to suboptimal representation learning for the under-represented classes.



(a) Balanced Pre-training (b) Tailed Drift Pre-training

Figure 1: The t-SNE visualization of feature space under the different conditions of pre-training within ImageNet and ImageNet-LT. The dark colors signify the region corresponding to the tail category with limited pre-training samples, whereas light colors denote the head category characterized by abundant samples.

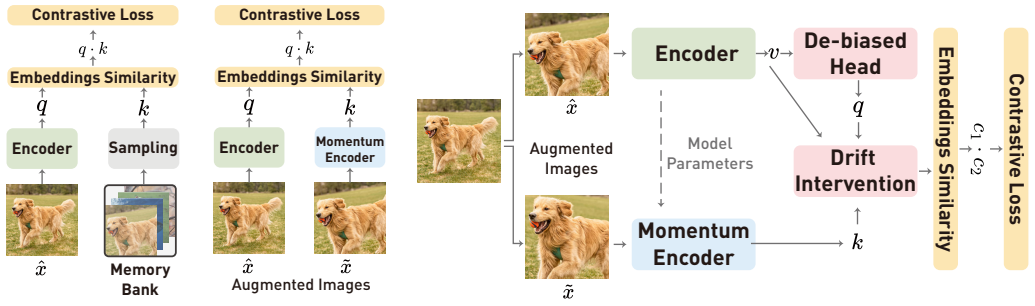
Building the above analysis, we formalize the interplay among these components using a Structural Causal Model (SCM), illustrated in Figure 3a. It elucidates the causal relationships among input samples X, predictions Y, the latent concept drift D inherent in the data stream, and the bias B accumulated in the momentum-updated teacher network (Pearl, 2014). Our causal analysis (detailed in Section 2.1) reveals that concept drift D acts as a critical confounder. It not only influences the input features X ($D \rightarrow X$) but also shapes the bias B within the teacher network $(X, D) \rightarrow B$, which in turn affects Y, i.e., $(X, B) \rightarrow Y$. This induces a spurious backdoor path $X \leftarrow D \rightarrow B \rightarrow Y$ that can mislead the learning process by inducing false correlations between X and Y. Furthermore, the mediation (Pearl, 2022) path $X \rightarrow B \rightarrow Y$ confounds the direct contribution of $X \rightarrow Y$, as detailed in Section 2.1.

These causal insights underscore the limitations of standard contrastive methods in drifting environments. Therefore, to counteract the detrimental effects of confounding and biased mediation, we propose a novel pre-training method, **R**esilient **C**ontrastive **P**re-training (**RCP**). RCP leverages causal intervention techniques (Pearl, 2016) to actively mitigate the bias B propagated by concept drift D, thereby enabling more robust and generalizable representation learning from non-stationary data. In summary, our paper mainly makes the following contributions:

1. We are pioneers in uncovering how concept drift impacts contrastive pre-training, specifically identifying the accumulation of bias through the momentum update mechanism. This foundational understanding paves the way for future research into robust pre-training on large-scale, dynamic datasets.

2. Leveraging insights from our structural causal model, we propose resilient contrastive pre-training (RCP), a novel method that employs causal intervention to mitigate drift-induced bias. RCP is designed to be straightforward to implement and scalable, making it suitable for pre-training large models on evolving data.
3. Extensive experiments demonstrate the superiority of RCP across diverse downstream tasks: long-tailed classification, OOD detection, and domain adaptation, using both fine-tuning and linear probing evaluation protocols. Notably, our experiments confirms that RCP effectively mitigates the detrimental effects of drift, leading to more discriminative and robust representations for large-scale pre-training.

2 METHODOLOGY



(a) Memory Bank (b) MoCo (c) Our Resilient Contrastive Learning

Figure 2: **Comparison between different contrastive pre-training paradigms.** (a) The key representations are sampled from a memory bank. (b) A momentum-updated encoder maintains the queue of keys. (c) The workflow of our resilient contrastive pre-training under concept drift streaming. Within the data streaming, a large batch size is opted for a wider drift adaptation window sliding to adapt changes in data distribution. Undergoes various random augmentations, the transformed instances from the identical sample are feature-extracted by both the encoder and the momentum encoder to get the key and value, respectively. A head is utilized to debias the drift and obtain the query of the encoder features. Subsequently, causal intervention is utilized to alleviate concept drift in the data stream within the adaptation window, resulting in the acquisition of two objects for contrastive learning.

2.1 CAUSAL LENS ON CONTRASTIVE DRIFT PRE-TRAINING

Contrastive Drift Pre-training. Concept drift is a statistical phenomenon wherein the joint distribution $P(X, Y)$ of features X and target variables Y evolves over time (Lu et al., 2019). We define Drift Pre-training as a pre-training paradigm operating on a data stream $\mathcal{S}_{0,T} = \{S_0, S_1, \ldots, S_T\}$. Each element $S_t = \{(x_i, y_i)\}_{i=1}^{n_t}$ consists of n_t samples with feature vector x_i and an associated contrastive label y_i at timestamp t. This timestamp t corresponds to an iteration in the pre-training process. At any given iteration t, data samples $\{X, Y\} \in \{(x_i, y_i)\}_{i=1}^{n_t}$ are drawn from a joint distribution $P_t(X, Y)$. Concept drift within this pre-training context is then formally characterized by:

$$\exists t \in [0, T): P_t(X, Y) \neq P_{t+1}(X, Y), \tag{1}$$

where the joint probability $P_t(X, Y)$ can be decomposed as $P_t(X, Y) = P_t(X) \times P_t(Y|X)$. Specifically, tailed drift (or long-tailed imbalance) primarily reflects variations in the marginal distribution $P_t(X)$, where the frequency of certain classes or patterns changes over time. In contrast, out-of-distribution (OOD) drift corresponds to alterations in the conditional distribution $P_t(Y|X)$, indicating that the mapping from features to labels evolves with new or redefined concepts. Both phenomena are manifestations of concept drift, where either the data distribution or the underlying labeling function (or both) changes unpredictably over time. In this work, we aim to develop a pre-training paradigm that can explicitly account for and adapt to these coupled forms of drift, enabling contrastive models to remain robust and consistent across dynamic environments.

Contrastive drift pre-training investigates contrastive learning under such non-stationary conditions, typically employing momentum-updated teacher-student frameworks (He et al., 2020), i.e.,

$$\theta_t^m = \lambda\theta_{t-1}^m + (1-\lambda)\theta_{t-1}^g, \tag{2}$$

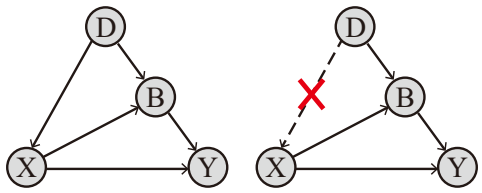
where θ^m signifies the parameters of the teacher network, θ^g denotes those of the student network, and $\lambda \in [0, 1)$ represents the momentum coefficient. Momentum $\lambda\theta_{t-1}^m$ induces a slower evolution of the teacher network θ^m relative to the student network θ^s. While high momentum (e.g., $\lambda = 0.999$ (Chen et al., 2021)) stabilizes training on stationary data, it poses a significant challenge under concept drift: the slowly updating teacher network θ^m can become severely misaligned with the current data distribution $P_t(x, z)$, thereby accumulating stale biases from past distributions and hindering the model's ability to adapt to new or evolved concepts present in the current iteration S_t.

Objective. Therefore, Contrastive drift pre-training aims to enable the learning system to synchronize with the evolving data distribution, ensuring that the pre-training model accurately reflect the current state of the data. We formalize this objective as maximizing the agreement between the student model and momentum model over a prospective adaptation window $[t, t + \tau]$:

$$\max_{\{g_j^\theta\}_{j=t}^{t+\tau}} \sum_{j=t}^{t+\tau} \mathcal{L}_{\text{agree}}(g_j^\theta(\tilde{X}_j), m_j^\theta(\hat{X}_j)) \tag{3}$$

where $\tilde{X}_j$ and $\hat{X}_j$ symbolize different augmented samples obtained from X_j, g_j^θ denotes the student model trained by the data stream $S_{j-\tau,j-1}$ from the drift adaption window with the size of τ, and m_j^θ represents the momentum model. And the pre-training is driven by the target metric $\mathcal{L}_{\text{agree}}$ to continuously adapt the drift in a given drift adaptation window of $[t, t + \tau]$ time period. This is a standard objective for continuous contrastive learning but fails to explicitly handle the accumulated bias B caused by drift D, which we formalize next.

Causal Analysis of Contrastive Drift Pre-training. To further understand the challenges posed by concept drift in contrastive drift pre-training, we employ a Structural Causal Model (SCM) (Pearl, 1995; 2016). As shown in Figure 3, the constructed causal graph of $\{X, Y, D, B\}$ presents the following causal connections:



(a) SCM (b) Intervention

Figure 3: The proposed causal graph of contrastive pre-training. **X**: Sample Features, **Y**: Prediction, **D**: Latent Concept Drift within Data, and **B**: Sample Bias in the Momentum Update.

$D \to X$: X denotes features extracted by samples drawn from the data stream with concept drift D, which is obviously trained under the effect of the drift pre-training.

$(X, D) \to B$: B represents the sample bias deviated from feature X under the effect of the concept drift D within the data stream. In the context of drift pre-training, the bias will accumulate during the momentum update of the teacher network, which will be amplified in the subsequent iteration of the contrastive pre-training.

$(X, B) \to Y$: This link presents that, apart from the regular $X \to Y$, the prediction is also impacted by the concept drift within the data stream through the mediation bias of B.

This structure reveals two detrimental pathways that compromise the true causal link $X \to Y$:

Backdoor Path (Spurious Correlation): $X \leftarrow D \to B \to Y$. Concept drift D acts as a confounder. This path introduces a spurious correlation, meaning X and Y appear related solely due to their mutual dependency on D and B. For instance, in long-tailed drift, tail samples might be spuriously associated with head categories due to the overwhelming head-class bias B.

Mediation Path (Weakening Effect): $X \to B \to Y$. The bias B acts as a mediator. This path conveys the drift effect, effectively capturing and weakening the direct and desirable impact of the true feature X on the prediction Y.

2.2 CAUSAL INTERVENTIONAL CONTRASTIVE OBJECTIVE

To eliminate the confounding influence of drift D (the backdoor path) and recover the genuine causal effect $X \rightarrow Y$, we employ a causal intervention. Drawing from the aforementioned causal relationships, it can be argued that current contrastive pertaining approaches involving the InfoNCE loss (Oord et al., 2018) essentially leverages the likelihood $P(Y|X)$ to drive the whole network, which is vulnerable by the concept drift confounder D and leads to the emergence of spurious correlations. By marginalizing over the discrete drift contexts $D = \{d_1, \ldots, d_{|D|}\}$ (Tang et al., 2021), $P_t(Y|X)$ can be expressed as:

$$P_t(Y|X) = \sum_{i=1}^{|D|} P_t(Y|X, B = h(X, d_i)) P_t(d_i|X) \tag{4}$$

where $B = h(X, d)$ represents the sample bias B conditioned on features X and a specific drift context d_i. Following the (Deng & Zhang, 2021), d_i is simplified to the drift of the latent category center underlying the pre-training. The term $P_t(d|X)$ highlights how X might carry information about d.

Inspired by recent advancements in applying causal inference to deep learning (Lv et al., 2022; Choi et al., 2022; Rohekar et al., 2023), we employ causal intervention $P_t(Y|\text{do}(X))$ to cut off the pathway of $D \rightarrow X$ at timestamp t in drift data stream as illustrated in Fig. 3b, where $\text{do}(\cdot)$ denotes the interventional operation (Pearl, 2016; Pearl et al., 2000). In accordance with the backdoor adjustment, the de-confounding operation $\text{do}(X)$ involves stratifying the confounder into discrete segments $D = \{d_1, d_2, \ldots, d_{|D|}\}$, thereby rendering D no longer a confounder between X and Y. Thus, $P_t(Y|\text{do}(X))$ can be expressed as:

$$P_t(Y|\text{do}(X)) = \sum_{i=1}^{|D|} P_t(Y|X, B = h(X, d_i)) P_t(d_i) \tag{5}$$

In Eq. equation 5, $P_t(Y|\text{do}(X))$ compels X to equally consider each confounding factor d_i and integrates them collectively to predict Y. Then, the original classifier $P_t(Y|X)$ is replaced by $P_t(Y|\text{do}(X))$ at the iteration t, as shown:

$$P_t(Y|\text{do}(X)) \triangleq P_t(Y|X) \tag{6}$$

The de-confound classifier mitigates the confounding influence and captures the genuine causality from X to Y, thereby enhancing the quality of the contrastive pre-training. Nevertheless, Eq. equation 5 necessitates costly sampling to approximate $P(Y|\text{do}(X))$ when implementing it in contrastive pre-training, resulting in impractical training times. Fortunately, inspired by prior studies that approximate Eq. equation 5 in a feed-forward manner, we follow this line of work and employ a single-pass estimation to approximate the intervention effect:

$$P_t(Y|\text{do}(X)) \approx P_t(Y|X, B = \sum_{i=1}^{|D|} h(X, d_i) P_t(d_i)) \tag{7}$$

As a result, according to the interventional probability outlined in Eq. equation 7, contrastive pre-training is compelled to grasp the genuine causal effect: $X \rightarrow Y$ rather than the misleading correlations induced by the concept drift confounder D. Thus, combined with Eq. equation 3, the final interventional contrastive pre-training objective is formulated as:

$$\begin{aligned} \mathcal{L}_{\text{agree}}(g_j^\theta(\tilde{X}_j), m_j^\theta(\hat{X}_j)) &= P_t(Y|X, \sum_{i=1}^{|D|} h(X, d_i) P_t(d_i)) \\ &= P_t(Y|g(\tilde{X}_j), \text{Softmax}(h(g_j^\theta(\tilde{X}_j m_j^\theta(\hat{X}_j))))), \end{aligned} \tag{8}$$

where the output of the final expression serves as the similarity score $\mathcal{L}_{\text{agree}}$ for the contrastive loss.

2.3 RESILIENT CONTRASTIVE PRE-TRAINING

This section details the concrete architecture and mechanism used to instantiate the causal objective (Eq. equation 8).

To realize the causal interventional objective for contrastive drift pre-training, we introduce **Resilient Contrastive Pre-training (RCP)**, an approach tailored for learning from data exhibiting concept drift. As depicted in Fig. 2, RCP adapts the student-teacher architecture of MoCo v3 (Chen et al., 2021), utilizing an encoder (g) and a momentum encoder (m). Within each drift adaptation window $[t, t+\tau]$, two augmented views ($\tilde{x}$, $\hat{x}$) are generated per image. The student encoder g processes one view into features $v = g(\tilde{x})$, which an MLP head then projects into a query $q = h(v)$. The momentum encoder m, structurally identical to g and updated via EMA (Eq. equation 2), processes the other view to produce a key $k = m(\hat{x})$, ensuring stable targets and enhancing training stability. Central to RCP is an intervention module that processes q, v, and k to mitigate latent biases from concept drift. This module computes an interventional score component $c = \text{Softmax}(q \cdot k) \cdot v$, which embodies aspects of our causal objective (Eq. equation 8). The overall network is then trained using an InfoNCE loss (Oord et al., 2018):

$$\mathcal{L} = -\log \frac{\exp\left(c_1 \cdot c_2^{i+}/\tau\right)}{\sum_{i=t}^{t+\tau} \exp\left(c_1 \cdot c_2^{i}/\tau\right)}, \tag{9}$$

where τ is a temperature hyper-parameter (Wu et al., 2018). The sum is over one positive and all negative samples within the drift adaptation window $[t, t+\tau]$ of the data stream.

Intuitively, q akin to a "query" (He et al., 2020), the momentum output k acts as the "key" for sampling key drift of the data stream, and the "value" v resembles a blend of the essential contrastive object and the concept drift. Therefore, the goal of causal intervention is to mitigate the bias of the contrastive objective through the drift sampling within the adaption window. We find the derived causal intervention module has a similar structure to the self-attention mechanism (Vaswani et al., 2017). Thus, this query-key-value mechanism structurally resembles self-attention (Vaswani et al., 2017), suggesting an inherent capability to model contextual dependencies useful for tracking drift. However, it is important to highlight that we employ "key" and "query" sampling for concept drift within the data stream to mitigate bias in the "value", which is different from the aim of sequence position modelling in the self-attention mechanism. Furthermore, our RCP method is simple to implement, efficient and scalable, with more details provided in Appendix.

3 Experiments

In this section, we initially showcase the robust performance of our RCP in the downstream task of long-tailed classification under drift pre-training, with results of fine-tuning and linear probing. Subsequently, from the perspective of generalization, we conduct experiments of out-of-distribution (OOD) detection and domain shift, to assess the resilience of the model against gradual drift and sudden drift. Following that, the feature embedding space of pre-trained models is intuitively visualized with degrees, illustrating how RCP mitigates the concept drift within the feature space. Finally, we present the scaling ability of our method facing with the drift pre-training.

It is worth noting that we do not assume any type of drift. We chose several benchmarks with known drifts, such as long tail, OOD, and domain shift, only for comparison with existing benchmarks, and visually illustrated the effectiveness of our model. However, we did not optimize for specific drift types during training. More detailed experimental implementations are given in Appendix.

3.1 Navigating Tailed Drift Pre-training

We compare our proposed RCP with other models to explicitly demonstrate its superior performance in tailed drift pre-training. Two large-scale datasets, ImageNet-LT (Liu et al., 2019) and iNaturalist 2018 (Van Horn et al., 2018b), are utilized as source datasets to perform pre-training, fine-tuning and linear probing, respectively. To effectively illustrate the efficacy of RCP in mitigating tail drift, we follow the criterion and metrics of long-tailed classification (Yang et al., 2023a) to report the Top-1 accuracy across various splits, including Many split with over 100 training samples, Medium split with 20-100 training samples, and Few split with fewer than 20 training samples.

In terms of fine-tuning results shown in Table 1, we compare mainstream contrastive pre-training methods, namely DeiT (Touvron et al., 2022), BYOL (Grill et al., 2020) , DINO (Caron et al., 2021) , DINO v2 (Zhou et al., 2024), MoCo v3 (Chen et al., 2021) with the baseline of ViT (Dosovitskiy et al., 2021), and our RCP exhibits superior results beyond other contrastive pre-training methods.

Table 1: Evaluation results of fine-tuning on long-tailed classification tasks with ImageNet-LT (Liu et al., 2019) and iNatualist2018 (Van Horn et al., 2018b). The best-performing contrastive pre-training results are highlighted in red. Many, Medium and Few denote the evaluated splits of many-shot (>100 training samples), medium-shot (20-100 samples) and few-shot (<20 samples). Top-1 accuracy is applied to evaluate the performance of different methods. † denotes methods that are adjusted for the long-tailed fine-tuning.

Methods	Venue	Backbones	ImageNet-LT				iNaturalist2018			
			Many	Medium	Few	All	Many	Medium	Few	All
Training from scratch										
cRT (Kang et al., 2019)	ICLR'19	ResNet-50	61.8	46.2	27.3	49.6	69.0	66.0	63.2	65.2
LWS (Kang et al., 2019)	ICLR'19		60.2	47.2	30.3	49.9	65.0	66.3	65.5	65.9
RIDE (Wang et al., 2020)	ICLR'20		68.2	53.8	36.0	56.9	70.9	72.4	73.1	72.6
PaCo (Cui et al., 2021)	ICCV'21		68.2	58.7	41.0	60.0	70.3	73.2	73.6	73.2
Masked Image Modeling Pre-training										
MAE (He et al., 2022)	CVPR'22	ViT-B/16	74.7	48.2	19.4	54.5	79.6	70.8	65.0	69.4
LiVT † (Xu et al., 2023)	CVPR'23		73.6	56.4	41.0	60.9	78.9	76.5	74.8	76.1
Contrastive Learning Pre-training										
BYOL (Grill et al., 2020)	NeurIPS'20	ViT-B/16	30.6	8.6	0.9	15.9	28.4	22.6	24.5	24.0
ViT (Dosovitskiy et al., 2021)	NeurIPS'21		50.5	23.5	6.9	31.6	65.4	55.3	50.9	54.6
DINO (Caron et al., 2021)	ICCV'21		64.8	35.5	11.3	43.3	76.7	67.1	61.2	65.5
DeiT (Touvron et al., 2022)	ECCV'22		70.4	40.9	12.8	48.4	72.9	62.8	55.8	61.0
MoCo v3 (Chen et al., 2021)	ICCV'21		70.8	40.7	14.3	48.5	76.6	65.8	62.8	65.6
DINO v2 (Zhou et al., 2024)	ICML'25		68.4	36.9	11.7	45.4	71.9	62.7	58.2	61.6
Ours			72.2	43.1	16.0	50.3	77.7	68.7	63.8	67.5

Compared to other methods that employ the student-teacher paradigm, such as BYOL, DINO, DINO v2, and MoCo v3, our RCP method demonstrates additional performance enhancements on both the Medium and Few splits, especially in the iNaturalist2018 dataset. It is demonstrated that the proposed causal interventional objective significantly alleviates the accumulated bias due to momentum updates of the teacher network. Meanwhile, under the concept drift scenario, BYOL without negative sample pairs has the inferior performance of long-tailed classification to the baseline ViT, while it is not noticeable in equilibrium (Chen et al., 2021). We argue that, in drift pre-training, the essence of negative sample pairs is similar to the causal interventional objective, taking into account the causal relationship of all drifts for each positive sample like de-confounding operation in Eq. equation 5 to delineate the boundary of the feature space. Relying solely on maximizing positive agreement as a pre-training strategy will cause the model to be overwhelmed by the head category, eventually exacerbating concept drift.

Besides, we also provide other fine-tuning results of long-tailed classification under scratch training and masked image modelling pre-training for comprehensive analysis, as shown in Table 1. It is worth noting that our focus is on the impact of drift environments on contrastive pre-training. Therefore, we exclusively utilized RCP solely during the pre-training phase, without making any adjustments to the downstream classification tasks. Consequently, in comparison to other methods, all contrastive pre-training approaches exhibit relatively lower performance. In the context of the masked image modeling strategy, it is found that contrastive learning is more susceptible to concept drift in the data stream. We attribute it to the fact that contrastive learning relies more heavily on global information within the data flow to construct feature representations.

Table 2: Evaluation results of linear probing on long-tailed classification tasks with ImageNet-LT and iNatualist2018. The best-performing models are highlighted in red.

Methods	Many	Medium	Few	All
ImageNet-LT (Liu et al., 2019)				
BYOL (Grill et al., 2020)	8.5	0.5	0.1	3.5
DINO (Caron et al., 2021)	47.4	20.5	0.1	27.9
DINO v2 (Zhou et al., 2024)	49.4	23.4	5.7	30.8
MoCo v3 (Chen et al., 2021)	45.8	18.7	0.0	27.1
ReLIC (Mitrovic et al., 2021)	3.2	21.0	47.1	28.6
Ours	52.7	24.0	7.9	32.8
iNaturalist2018 (Van Horn et al., 2018b)				
BYOL (Grill et al., 2020)	7.0	3.1	2.5	3.2
DINO (Caron et al., 2021)	43.1	33.5	32.3	34.0
DINO v2 (Zhou et al., 2024)	36.3	30.2	30.0	30.7
MoCo v3 (Chen et al., 2021)	32.9	25.4	22.4	24.9
Ours	48.4	37.0	35.1	37.3

Furthermore, we provide linear probing results of contrastive pre-training methods in Table 2. It is evidenced that our method outperforms other contrastive pre-training methods, corroborating our main contributions of causal interventional contrastive objective for drift pre-training. It is noteworthy that on the Few split of ImageNet-LT, several contrastive learning methods failed to perform effectively, whereas we attained an accuracy of 7.9%. It indicates that our approach can accurately construct feature representations for tail categories even in the presence of concept drift.

3.2 Taming ID and OOD Drift

To validate the generalization capability of RCP under drifting scenarios, we conducted experiments from the perspectives of domain shift and out-of-distribution (OOD) detection.

In terms of domain shift generalization as exhibited in Table 3, our primary focus is to assess the extent of the drift environment on the feature representation constructed during pre-training. Therefore, we select three subsets of ImageNet (Russakovsky et al., 2015) to validate the domain shift performance, namely, ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019) and ImageNet-R (Hendrycks et al., 2021a). Our results achieve superior results across three datasets, investigating the effectiveness of the proposed RCP method. It corroborates that, the feature space we construct in pre-training abstracts and captures the essential information of images, while disregarding any unexpected interference from the concept drift within the data stream. Moreover, judging from the results of Few split in ImageNet-Sketch and ImageNet-R, the majority of contrastive pre-training methods struggle to differentiate the tail categories. This difficulty arises from two main challenges. First, the bias induced by tail drift tends to skew models toward favoring head categories. On the flip side, the small sample size of tail categories limits the ability of pre-training methods to accurately extract their fundamental features, a problem that is exacerbated in the presence of domain shifts. Accordingly, our outstanding results demonstrate our proficiency in mitigating tail drift and extracting crucial features from limited samples.

Table 3: Evaluation results of domain shift on ImageNet-V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019) and ImageNet-R (Hendrycks et al., 2021a). The best-performing models are highlighted in red. Top-1/-5 accuracy are applied to evaluate the performance.

Methods	Many	Medium	Few	Top-1	Top-5
ImageNet-V2 (Recht et al., 2019)					
BYOL (Grill et al., 2020)	29.7	7.8	1.0	15.2	31.8
DINO (Caron et al., 2021)	63.4	33.6	11.6	41.9	65.1
DINO v2 (Zhou et al., 2024)	66.6	35.4	11.9	44.0	66.9
MoCo v3 (Chen et al., 2021)	62.9	37.3	12.6	43.1	66.0
Ours	69.7	41.2	14.9	48.3	72.1
ImageNet-Sketch (Wang et al., 2019)					
BYOL (Grill et al., 2020)	1.9	0.2	0.0	0.8	2.8
DINO (Caron et al., 2021)	12.4	5.0	0.5	7.1	16.2
DINO v2 (Zhou et al., 2024)	17.1	6.8	1.4	9.9	20.8
MoCo v3 (Chen et al., 2021)	17.6	7.6	1.4	10.4	21.9
Ours	22.2	9.7	2.3	13.4	27.0
ImageNet-R (Hendrycks et al., 2021a)					
BYOL (Grill et al., 2020)	4.2	1.7	0.0	2.7	6.8
DINO (Caron et al., 2021)	13.4	7.6	0.4	9.5	18.3
DINO v2 (Zhou et al., 2024)	15.8	8.6	0.7	11.1	21.1
MoCo v3 (Chen et al., 2021)	17.2	9.8	1.0	12.3	22.8
Ours	21.3	12.5	1.6	15.5	27.5

Concerning out-of-distribution (OOD) detection, we also evaluate the capability of our RCP approach to delineate the boundaries of the feature space, as presented in Table 4. While exhibiting slightly lower performance than DINO on ImageNet-O, we significantly outperform contrastive pre-training methods on the other three out-of-distribution datasets, and surpasses other approaches in overall performance. It demonstrates that our model possesses the benefits of intra-class compactness and inter-class separability in feature representation. Meanwhile, it is worth noting that in larger-scale balanced pre-training datasets, such as ImageNet-21k (Ridnik et al., 2021), our approach outperforms numerous traditional OOD detection methods in overall performance, such as ViM (Wang et al., 2022a). It underscores our RCP method can effectively alleviate concept drift bias in the model while also enhancing the characterization of the model's decision boundary.

3.3 Boosting Pre-training Feature Embedding

To directly demonstrate the feature space of the pre-trained model, we quantitatively calculate the feature embedding distances from three perspectives: In-Distribution (ID) intra-class compactness,

Table 4: Evaluation results of OOD detection with the OOD datasets of Texture (Cimpoi et al., 2014), iNat-OOD (Van Horn et al., 2018a; Huang & Li, 2021), ImageNet-O (Hendrycks et al., 2021b) and OpenImage-O (Wang et al., 2022b). ViT-B/16 is selected as the image encoder. The best-performing method is highlighted in red. FPR↓ and AUROC↑ are applied to evaluate the performance of different methods. PT denotes the pre-training dataset, while ID means the in-distribution dataset.

Methods	Texture		iNaturalist		ImageNet-O		OpenImage-O		Overall	
	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR	AUROC	FPR
PT: ImageNet-21k, ID: ImageNet-1K										
Residual (Wang et al., 2022a)	82.4	64.6	73.7	86.0	68.4	87.5	74.9	78.0	74.9	79.0
React (Sun et al., 2021)	62.1	80.5	91.2	38.7	63.7	81.0	80.4	60.4	74.3	65.2
ViM (Wang et al., 2022a)	83.5	62.7	77.8	81.7	71.0	86.6	78.3	74.6	77.7	76.4
MOODv1 (Li et al., 2023)	93.0	30.9	98.8	5.9	86.8	63.2	95.5	26.5	93.5	31.6
MOODv2 (Li et al., 2024)	94.3	24.7	99.6	1.8	91.5	40.8	97.4	13.6	95.7	20.2
PT: ImageNet-LT, ID: ImageNet-LT										
BYOL (Grill et al., 2020)	58.4	92.1	80.1	77.5	52.0	92.2	65.3	81.3	64.0	85.8
DINO (Caron et al., 2021)	72.9	80.7	85.7	65.3	74.2	80.2	74.7	83.0	76.9	77.3
DINO v2 (Zhou et al., 2024)	67.1	85.1	78.9	88.4	63.1	88.3	73.9	85.4	70.7	86.8
MoCo v3 (Chen et al., 2021)	81.5	63.9	80.1	68.6	65.0	91.0	69.9	84.2	74.1	76.9
Ours	83.1	57.8	88.0	67.7	70.1	82.9	75.5	78.3	79.2	71.7

ID inter-class separability, and the separability between ID and OOD categories, as shown in Table 5. Intra-class compactness within the In-Distribution (ID) measures the average distance between the category center and samples. ID Inter-class separability evaluates the distances between centers of different categories. Lastly, we assess the separability between ID categories and OOD samples.

The intra-class compactness results demonstrate that our model effectively validates the efficacy of the proposed RCP in delineating feature boundaries, especially in long-tailed scenarios. Moreover, the inter-class separability among different splits is very similar, suggesting that the pre-training primarily influences intra-class compactness rather than inter-class separability. Furthermore, with the exception of BYOL, the inter-class separability of the other contrastive pre-training methods is nearly the same. This observation underscores the significance of negative samples in expanding inter-class separation. It also demonstrates that the bias is induced by accumulated drift within momentum updates, where features of head categories dominated the whole model under tailed drift. In the context of ID vs. OOD separability, we attained the optimal results, signifying that our feature space exhibits distinct feature boundaries. In addition, we find that the performance of ID vs. OOD separability aligns with the ID intra-class compactness, suggesting that enhancing the model's performance hinges on mitigating data stream drift to enhance feature extraction capabilities.

Table 5: Evaluation results of different contrastive learning strategies in pre-training from three perspectives: ID intra-class compactness, ID inter-class separability and the separability between ID and OOD categories. The cosine metric is utilized to measure these distances, which is expressed as average degrees. ImageNet-LT is utilized for pre-training. Best results are highlighted in red.

Pre-training	ID Intra Compactness ↓				ID Inter Separability ↑				ID vs. OOD Separability ↑			
	Many	Medium	Few	All	Many	Medium	Few	All	Many	Medium	Few	All
BYOL (Grill et al., 2020)	53.1	59.1	59.9	56.9	81.5	79.2	78.3	80.0	75.7	71.9	70.3	73.1
DINO (Caron et al., 2021)	34.2	51.4	59.0	45.9	89.0	88.4	87.6	88.5	86.2	84.4	81.4	84.7
DINO v2 (Zhou et al., 2024)	31.2	50.1	57.8	43.9	88.9	88.1	87.2	88.3	86.0	83.7	80.5	84.1
MoCo v3 (Chen et al., 2021)	29.1	47.8	57.8	42.8	89.3	88.8	88.2	88.9	86.9	85.7	83.0	85.8
Ours	28.5	46.5	55.8	40.9	89.4	89.0	88.4	89.1	87.1	86.1	83.4	86.1

4 Conclusion and Outlooks

In this paper, we present resilient contrastive pre-training (RCP), a novel, straightforward and effective pre-training paradigm tailored for concept drift data stream. We employ causal inference theory to methodically examine the source of bias in the momentum update of contrastive pre-training and put forward a causal interventional objective to mitigate this bias within the drifting data stream. By virtue of this objective, RCP is devised to counteract the unpredictable distribution changes occurring within the data stream.

AI USE STATEMENT

This work has not used generative AI tools for research ideation, experimental design, data analysis, or result generation. The authors take responsibility for the final content of this work.

ETHICS STATEMENT

This work studies representation learning under non-stationary data. We do not introduce new datasets involving human subjects. Experiments use standard public benchmarks, and we do not anticipate direct ethical risks beyond those commonly associated with representation learning models.

REPRODUCIBILITY STATEMENT

We provide implementation details, model settings, datasets, and evaluation protocols in the main paper. The anonymized code repository is provided in the abstract and will be maintained with instructions for reproducing the reported experiments.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.

Vitor Cerqueira, Heitor Murilo Gomes, Albert Bifet, and Luis Torgo. STUDD: A student–teacher method for unsupervised concept drift detection. 112(11):4351–4378, 2023. ISSN 1573-0565. doi: 10.1007/s10994-022-06188-7. URL https://doi.org/10.1007/s10994-022-06188-7.

Hong-You Chen, Cheng-Hao Tu, Ziwei Li, Han Wei Shen, and Wei-Lun Chao. On the importance and applicability of pre-training for federated learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9640–9649, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_An_Empirical_Study_of_Training_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Yufei Chen, Xiaoyu Yang, Xiaodong Yue, Xiang Lin, Qi Zhang, and Hamido Fujita. A general variation-driven network for medical image synthesis. *Applied Intelligence*, 54(4):3295–3307, 2024.

Seungtaek Choi, Myeongho Jeong, Hojae Han, and Seung-won Hwang. C2L: Causally Contrastive Learning for Robust Text Classification. 36(10):10526–10534, 2022. ISSN 2374-3468. doi: 10.1609/aaai.v36i10.21296. URL https://ojs.aaai.org/index.php/AAAI/article/view/21296.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 715–724, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Cui_Parametric_Contrastive_Learning_ICCV_2021_paper.html.

Xiang Deng and Zhongfei Zhang. Comprehensive Knowledge Distillation with Causal Intervention. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22158–22170. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/b9f35816f460ab999cbc168c4da26ff3-Abstract.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. URL http://arxiv.org/abs/2010.11929.

Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10112–10121, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/He_Momentum_Contrast_for_Unsupervised_Visual_Representation_Learning_CVPR_2020_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. pp. 15262–15271, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Rui Huang and Yixuan Li. MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Huang_MOS_Towards_Scaling_Out-of-Distribution_Detection_for_Large_Semantic_Space_CVPR_2021_paper.html.

Meirui Jiang, Zirui Wang, and Qi Dou. Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1087–1095, 2022.

Botao Jiao, Yinan Guo, Dunwei Gong, and Qiuju Chen. Dynamic Ensemble Selection for Imbalanced Data Streams With Concept Drift. 35(1):1278–1291, 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2022.3183120. URL https://ieeexplore.ieee.org/abstract/document/9802893.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling Representation and Classifier for Long-Tailed Recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=r1gRTCVFvB.

Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11578–11589, 2023.

Jingyao Li, Pengguang Chen, Shaozuo Yu, Shu Liu, and Jiaya Jia. MOODv2: Masked Image Modeling for Out-of-Distribution Detection. 46(12):8994–9003, 2024. ISSN 0162-8828. doi: 10.1109/TPAMI.2024.3412004. URL https://www.computer.org/csdl/journal/tp/2024/12/10553645/1XH2Iao33Z6.

Wendi Li, Xiao Yang, Weiqing Liu, Yingce Xia, and Jiang Bian. DDG-DA: Data Distribution Generation for Predictable Concept Drift Adaptation. 36(4):4092–4100, 2022-06-28. ISSN 2374-3468. doi: 10.1609/aaai.v36i4.20327. URL https://ojs.aaai.org/index.php/AAAI/article/view/20327.

Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. *Advances in Neural Information Processing Systems*, 35:19645–19658, 2022.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, Deconfound and Tell: Image Captioning With Causal Inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18041–18050, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Show_Deconfound_and_Tell_Image_Captioning_With_Causal_Inference_CVPR_2022_paper.html.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2532–2541. IEEE, 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00264. URL https://ieeexplore.ieee.org/document/8953407/.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under Concept Drift: A Review. 31(12):2346–2363, 2019. ISSN 1558-2191. doi: 10.1109/TKDE.2018.2876857. URL https://ieeexplore.ieee.org/abstract/document/8496795.

Jie Lu, Anjin Liu, Yiliao Song, and Guangquan Zhang. Data-driven decision support under concept drift in streamed big data. *Complex & intelligent systems*, 6(1):157–163, 2020.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality Inspired Representation Learning for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8046–8056, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Lv_Causality_Inspired_Representation_Learning_for_Domain_Generalization_CVPR_2022_paper.html.

Qiaowei Miao, Junkun Yuan, and Kun Kuang. Domain Generalization via Contrastive Causal Learning, 2022. URL http://arxiv.org/abs/2210.02655.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Kunjal Panchal, Sunav Choudhary, Subrata Mitra, Koyel Mukherjee, Somdeb Sarkhel, Saayan Mitra, and Hui Guan. Flash: Concept drift adaptation in federated learning. In *International Conference on Machine Learning*, pp. 26931–26962. PMLR, 2023.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4): 459, 2014.

Judea Pearl. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pp. 373–392. 2022.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. ImageNet-21K Pretraining for the Masses, 2021. URL http://arxiv.org/abs/2104.10972.

Raanan Y. Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal Interpretation of Self-Attention in Pre-Trained Transformers. 36:31450–31465, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/642a321fba8a0f03765318e629cb93ea\-Abstract-Conference.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. 115(3):211–252, 2015.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. In *Advances in Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc//paper/2020/file/1091660f3dff84fd648efe31391c5524-Paper.pdf.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. DeiT III: Revenge of the ViT. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 516–533. Springer Nature Switzerland, 2022. ISBN 978-3-031-20053-3. doi: 10.1007/978-3-031-20053-3_30.

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*, pp. 8769–8778, 2018a.

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*, pp. 8769–8778, 2018b. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Van_Horn_The_INaturalist_Species_CVPR_2018_paper.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning Robust Global Representations by Penalizing Local Predictive Power. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/3eefceb8087e964f89c2d59e8a249915-Abstract.html.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022a.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-Distribution With Virtual-Logit Matching. pp. 4921–4930, 2022b. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_ViM_Out-of-Distribution_With_Virtual-Logit_Matching_CVPR_2022_paper.html.

Kun Wang, Li Xiong, Anjin Liu, Guangquan Zhang, and Jie Lu. A self-adaptive ensemble for user interest drift learning. 577:127308, 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.127308. URL https://www.sciencedirect.com/science/article/pii/S0925231224000791.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense Contrastive Learning for Self-Supervised Visual Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Dense_Contrastive_Learning_for_Self-Supervised_Visual_Pre-Training_CVPR_2021_paper.html.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=D9I3drBz4UC.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Wu_Unsupervised_Feature_Learning_CVPR_2018_paper.html.

Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. Learning Imbalanced Data With Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15793–15803, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Xu_Learning_Imbalanced_Data_With_Vision_Transformers_CVPR_2023_paper.html.

Xiaoyu Yang, Yufei Chen, Xiaodong Yue, Xiang Lin, and Qi Zhang. Variational synthesis network for generating micro computed tomography from cone beam computed tomography. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1611–1614. IEEE, 2021a.

Xiaoyu Yang, Yufei Chen, Xiaodong Yue, Chao Ma, and Panpan Yang. Local linear embedding based interpolation neural network in pancreatic tumor segmentation. *Applied Intelligence*, 52 (8):8746–8756, 2022.

Xiaoyu Yang, Yufei Chen, Xiaodong Yue, Shaoxun Xu, and Chao Ma. T-distributed Spherical Feature Representation for Imbalanced Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10825–10833, 2023a. ISSN 2374-3468. doi: 10.1609/aaai.v37i9.26284. URL https://ojs.aaai.org/index.php/AAAI/article/view/26284.

Xiaoyu Yang, Jie Lu, and En Yu. Adapting multi-modal large language model to concept drift from pre-training onwards. *arXiv preprint arXiv:2405.13459*, 2024a.

Xiaoyu Yang, Lijian Xu, Simon Yu, Qing Xia, Hongsheng Li, and Shaoting Zhang. Segmentation and vascular vectorization for coronary artery by geometry-based cascaded neural network. *IEEE Transactions on Medical Imaging*, 2024b.

Xiaoyu Yang, Jie Lu, and En Yu. Learning from all: Concept alignment for autonomous distillation from multiple drifting mllms. *arXiv preprint arXiv:2510.04142*, 2025a.

Xiaoyu Yang, Jie Lu, and En Yu. Walking the tightrope: Disentangling beneficial and detrimental drifts in non-stationary custom-tuning. *arXiv preprint arXiv:2505.13081*, 2025b.

Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal Attention for Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9847–9857, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_Causal_Attention_for_Vision-Language_Tasks_CVPR_2021_paper.html.

Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded Image Captioning: A Causal Retrospect. 45(11):12996–13010, 2023b. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3121705. URL https://ieeexplore.ieee.org/abstract/document/9583890.

En Yu, Yiliao Song, Guangquan Zhang, and Jie Lu. Learn-to-adapt: Concept drift adaptation for hybrid multiple streams. 496:121–130, 2022. ISSN 0925-2312. doi: 10.1016/j.neucom.2022.05.025. URL https://www.sciencedirect.com/science/article/pii/S0925231222005550.

En Yu, Jie Lu, Bin Zhang, and Guangquan Zhang. Online boosting adaptive learning under concept drift for multistream classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16522–16530, 2024.

En Yu, Jie Lu, Kun Wang, Xiaoyu Yang, and Guangquan Zhang. Drift-aware collaborative assistance mixture of experts for heterogeneous multistream learning. *arXiv preprint arXiv:2508.01598*, 2025.

Hang Yu, Weixu Liu, Jie Lu, Yimin Wen, Xiangfeng Luo, and Guangquan Zhang. Detecting group concept drift from multiple data streams. 134:109113, 2023. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.109113. URL https://www.sciencedirect.com/science/article/pii/S0031320322005933.

Congzhi Zhang, Linhai Zhang, and Deyu Zhou. Causal Walk: Debiasing Multi-Hop Fact Verification with Front-Door Adjustment, 2024. URL http://arxiv.org/abs/2403.02698.

Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *Proceedings of the 29th annual international conference on mobile computing and networking*, pp. 1–15, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning, 2024. URL http://arxiv.org/abs/2411.04983.

A Related Works

A.1 Concept Drift

In the comprehensive survey conducted by Lu et al. (Lu et al., 2019; 2020), existing approaches for concept drift handling are systematically classified into three primary categories: error rate-based methods (Wang et al., 2024; Jiao et al., 2024), data distribution-based methods (Yang et al., 2024a; Cerqueira et al., 2023), and multiple hypothesis-based techniques (Yu et al., 2024; 2022; 2025). Our proposed methodology falls under the category of distribution-based concept drift adaptation approaches. These distribution-driven techniques distinguish themselves by not only enabling precise drift identification through explicit statistical distribution analysis but also providing a multidimensional characterization of drift patterns - including temporal occurrence detection, affected feature space localization, and quantitative severity assessment. This dual capability of detection coupled with comprehensive drift diagnostics establishes distribution-based methods as particularly valuable for developing adaptive systems that require interpretable drift understanding and targeted model adjustments.

Besides, Online Boosting Adaptive Learning (OBAL) method (Yu et al., 2024) is proposed to address the challenges of concept drift and negative transfer in multistream classification. OBAL employs a dual-phase approach: an initial model is built using the Adaptive Covariate Shift Adaptation (AdaCOSA) algorithm to handle covariate shifts and learn dynamic correlations. In the online phase, a Gaussian Mixture Model-based weighting mechanism is integrated to manage asynchronous drift. Meanwhile, CDMLLM (Yang et al., 2024a) reveals that vision-language models suffer significant bias from concept drift during pre-training and fine-tuning. To address this, the authors propose a unified concept drift framework integrating T-distribution-based adaptation for long-tailed calibration and explicit OOD detection, demonstrating enhanced robustness in open-world multi-modal alignment through systematic distribution modelling. Beyond on drift detection in single data stream, GDDM (Yu et al., 2023) focuses on addressing group concept drift across multiple data streams, where individual drifts may go undetected due to subtle changes in underlying distributions. The proposed method introduces a distribution-free test statistic to detect concept drift in these complex scenarios. By designing an online learning algorithm for streaming data, the approach accurately identifies concept drift caused by hypothesis testing. Beyond that, DDG-DA (Li et al., 2022-06-28) proactively models predictable factors influencing environmental evolution. The approach involves training a predictor to estimate future data distribution, using this information to generate training samples, and then training models on the generated data. By leveraging predictable factors to forecast data distribution changes, DDG-DA aims to enhance model performance in handling concept drift in streaming data. Furthermore, STUDD(Cerqueira et al., 2023) proposes a teacher-student paradigm to enable unsupervised drift detection through deviation analysis of their predictive consistency. This approach leverages model disagreement as a proxy signal, bypassing dependency on ground-truth labels during deployment while maintaining detection sensitivity.

A.2 Causal Inference

Recently, increasing researchers have incorporated causal inference into deep-learning models, especially in large models. Deconfounded Image Captioning (DIC) (Yang et al., 2023b) is proposed to address dataset bias in vision-language models through a causal lens, that integrates backdoor and front-door adjustments for systematic bias mitigation. The framework provides principled causal analysis of spurious correlations in multimodal alignment, offering theoretical grounding for decomposing bias sources through structured interventions. Likewise, aiming for spurious correlations induced by visual and linguistic biases during training, CIIC (Liu et al., 2022) is proposed as a causal intervention framework combining an Interventional Object Detector (IOD) and Interventional Transformer Decoder (ITD) guided by structural causal models. By applying backdoor adjustment through IOD's feature disentanglement and ITD's dual de-confounding mechanism, their approach systematically mitigates confounding effects across encoding and decoding stages, demonstrating enhanced generalization through causal correlation modeling. Similarly, targeting multi-hop fact verification bias in the large language model, Causal Walk (Zhang et al., 2024) is proposed, a front-door adjustment framework that disentangles complex spurious correlations in evidence chains. The method models reasoning paths as mediators in structural causal models, decomposing causal effects via random walk-based treatment-mediator estimation and geometric mean-based

mediator-outcome approximation. By integrating adversarial and symmetric datasets synthesized with large language models, the approach demonstrates superior debiasing performance.

Additionally, causal inference is widely used in representation learning. Comprehensive Interventional Distillation (CID) (Deng & Zhang, 2021) integrates causal intervention with class-aware representation alignment. By reinterpreting teacher logits as contextual confounders and applying counterfactual pruning through structural causal models, CID systematically disentangles beneficial semantic patterns from dataset-specific biases. This approach demonstrates enhanced generalization through bias-invariant knowledge transfer. Besides, De-confound-TDE (Tang et al., 2021) establishes a causal framework for long-tailed classification, identifying SGD momentum as a paradoxical confounder that simultaneously harms tail-class predictions while benefiting representation learning. Through causal intervention during training and counterfactual reasoning at inference, the method disentangles momentum's detrimental bias from its beneficial mediation effects. Meanwhile, CCM (Miao et al., 2022) is proposed to address domain generalization through causal invariance principles. The framework integrates front-door adjustment with contrastive learning to quantify stable causal effects across domains, explicitly modeling domain shifts via a three-stage process: domain-conditioned supervision for feature correlation, causal effect measurement through structured path manipulation, and contrastive clustering for class-consistent representations. Similarly, CIRL (Lv et al., 2022) advances domain generalization through causal factorization, proposing a structural causal model that decomposes inputs into invariant causal mechanisms and domain-specific non-causal factors. It enforces three critical properties: causal/non-causal separation, joint independence, and causal sufficiency for classification.

Besides, C2L (Choi et al., 2022) addresses model fragility to spurious patterns through contrastive counterfactual synthesis, proposing a collective decision framework that aggregates predictions across generated counterfactual sets. Unlike conventional augmentation limited by dataset-inherent biases, this approach probabilistically supervises causal invariance through distributional consensus, demonstrating enhanced robustness against attribution bias and domain shifts. Furthermore, ABCD (Rohekar et al., 2023) establishes a causal interpretation of Transformer self-attention mechanisms, modeling them as structural equation estimators that capture conditional independence relations through partial correlation analysis in deep attention layers. This framework enables zero-shot causal discovery over input sequences while accounting for latent confounders, effectively repurposing pre-trained models for causal graph inference. What's more, Causal Attention (CATT) (Yang et al., 2021b) implements front-door adjustment to address confounding bias in attention mechanisms via dual-path processing of In-Sample and Cross-Sample Attention. By forcibly integrating external sample contexts through CS-ATT while maintaining standard attention conventions, CATT dynamically mitigates spurious correlations without requiring confounder specification.

A.3 Contrastive Pre-training

The seminal work of SimCLR (Chen et al., 2020) establishes foundational principles for contrastive pretraining through instance discrimination objectives and systematic augmentation strategies. While achieving remarkable performance in static environments, its reliance on fixed augmentation policies and stationary negative sampling assumes temporal consistency of latent patterns – an assumption violated under concept drift scenarios. Nevertheless, SimCLR's demonstration of invariant representation learning through normalized temperature-scaled loss provides crucial architectural groundwork for developing drift-aware contrastive frameworks, particularly in modeling evolving feature relationships through dynamic positive/negative pair formulation.

Besides, the MoCo framework (He et al., 2020) and its subsequent evolution MoCo v3 (Chen et al., 2021) establish critical momentum-based mechanisms for contrastive learning through dynamic memory banks and stabilized key encoder updates. It is primarily designed to maintain consistency in negative sample maintenance and gradient stability. However, their fixed momentum schedules and stationary target assumptions limit adaptability to abrupt distribution shifts as we discussed aforementioned. Likewise, BYOL (Grill et al., 2020) employs two neural networks—an online network and a target network—that interact and learn from each other. Specifically, the online network is trained to predict the representation of an augmented image view produced by the target network, while the target network is updated using a slow-moving average of the online network's parameters. Similarly, the SiamSim framework (Chen & He, 2021) advances contrastive representation

learning through dynamic similarity calibration, employing a dual-path Siamese architecture with momentum-aligned feature projectors.

Meanwhile, DINO (Caron et al., 2021) and DINO v2 (Zhou et al., 2024) advance self-distillation paradigms through momentum-based teacher-student mechanisms, leveraging global-local attention consistency in Vision Transformers to learn semantically structured representations. While achieving state-of-the-art in static self-supervised learning, their fixed teacher-student update schedules and monolithic prototype banks implicitly assume stationarity of feature distributions. DINO v2's introduction of partitioned expertise through specialized sub-networks demonstrates partial adaptability to data variations, suggesting pathways for concept drift mitigation through dynamic expert routing. These works collectively establish critical baselines for stability-aware distillation architectures in non-stationary pre-training scenarios.

Furthermore, DenseCL (Wang et al., 2021) advances contrastive learning through localized feature alignment, introducing dense instance discrimination that operates at both image-level and pixel-level granularity. By enforcing spatial consistency through region-to-region contrastive pretext tasks, the method enhances model sensitivity to fine-grained visual patterns. While primarily designed for dense prediction tasks, its hierarchical contrast mechanism demonstrates the importance of multi-scale feature stabilization.

While contemporary contrastive pretraining methods achieve remarkable performance under static data distributions, their reliance on closed-world stationarity assumptions presents a fundamental limitation: they inherently lack mechanisms for concept drift adaptation when deployed in non-stationary environments with evolving data.

B More Ablation Studies

B.1 Scaling with Drift Pre-training

Table 6: Evaluation results of scaling ability of our RCP on ImageNet-LT with various ViT models. Top-1 accuracy is utilized to evaluate the performance.

Backbones	Many	Medium	Few	All
ViT-Small/16	67.7	39.6	11.6	46.4
ViT-Base/16	72.2	43.1	16.0	50.3
ViT-Large/16	73.8	45.0	16.9	52.0

To present the scalability of our RCP approach in scenarios involving concept drift, we evaluate the efficiency and scalability of our model in Table 6. As the number of parameters increases, our model exhibits a scalable effect, leading to performance improvements with larger models in downstream classification tasks. In comparison to ViT-S/16, our proposed method achieves nearly a 4% enhancement with ViT-B/16. This suggests that we can effectively train larger models to attain superior performance in the face of concept drift within data.

More critically, our experiments reveal that pre-trained models exhibit consistent scaling behavior even on datasets with great distribution drift. It demonstrates that our proposed RCP can effectively leverage larger-scale, minimally curated datasets, thereby substantially expanding the pool of usable data for pre-training without relying on complicated cleaning procedures.

Table 7: Ablation experiments on adaptation window length within the linear probing task on ImageNet-LT. We also provide the results of MoCo v3 for comparison.

	Window	Many	Medium	Few	All
MoCo v3 (Chen et al., 2021)	4,096	45.8	18.7	0.0	27.1
Ours	2,048	42.1	12.8	1.1	22.4
Ours	4,096	47.9	20.1	3.2	28.5
Ours	9,600	52.7	24.0	7.9	32.8

B.2 Ablation on Adaptation Window Length

Furthermore, we also conduct ablation experiments on adaptation window length, as exhibited in Table 7. It demonstrates that our approach relies on a wider drift adaptation window, with superior results under a larger one. In particular, our method consistently outperforms MoCo v3 on Few Split, demonstrating the robustness of our method against non-stationary drift.