# Revisiting Supervision for Continual Representation Learning

**Daniel Marczak**[*1,2], **Sebastian Cygert**[1,3], **Tomasz Trzciński**[1,2,4], **Bartłomiej Twardowski**[1,5,6]
[1]IDEAS NCBR; [2]Warsaw University of Technology; [3]Gdańsk University of Technology;
[4]Tooploox; [5]Autonomous University of Barcelona; [6]Computer Vision Center

## Abstract

In the field of continual learning, models are designed to learn tasks one after the other. While most research has centered on supervised continual learning, recent studies have highlighted the strengths of self-supervised continual representation learning. The improved transferability of representations built with self-supervised methods is often associated with the role played by the multi-layer perceptron projector. In this work, we depart from this observation and reexamine the role of supervision in continual representation learning. We reckon that additional information, such as human annotations, should not deteriorate the quality of representations. Our findings show that supervised models when enhanced with a multi-layer perceptron head, can outperform self-supervised models in continual representation learning.

## 1 Introduction

In continual learning (CL), the goal of the model is to learn new tasks sequentially. A number of recent works study continual learning from a representation learning perspective and show that unsupervised approaches build more robust representations when trained continually: Madaan et al. (2022) shows that self-supervised learning (SSL) methods build representations that are more robust to forgetting than supervised learning (SL) while Davari et al. (2022) notices that training SimCLR (Chen et al., 2020) have advantageous properties for continual learning compared to SL. However, it is still counter-intuitive that access to more information (labels) results in worse representations in continual learning.

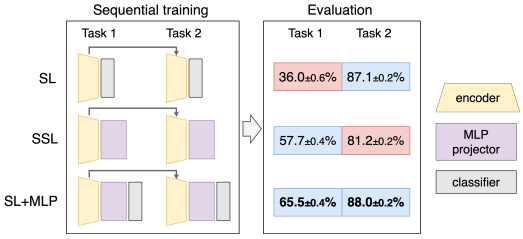


Figure 1: In a two-task continual learning scenario, supervised learning (SL) results in representations that perform well on the second task but poorly on the first task. On the other hand, representations trained with self-supervised learning (SSL) have higher first-task performance but they underperform on the second task. We show that simple modifications to supervised learning (SL+MLP) yield representations that are superior on the first task and on par with SL on the second task. We report average over 6 different scenarios.

One of the potential reasons is the transferability gap between supervised and unsupervised learning. It was believed that the superior transferability of unsupervised learning can be attributed to a special design of contrastive loss (Zhao et al., 2020) or lack of annotations during training (Ericsson et al., 2020). However, recent works (Wang et al., 2021; Sariyildiz et al., 2023) identify that a multi-layer perceptron (MLP) projector commonly used in SSL (Chen & He, 2020; Grill et al., 2020) is a crucial component that improves transferability

*Corresponding author, email: daniel.marczak.dokt@pw.edu.pl

Preprint.

of SSL models. Following that founding Wang et al. (2021); Sariyildiz et al. (2023) use an MLP projector to improve transferability of SL and achieve state-of-the-art transfer learning performance, surpassing unsupervised methods.

In this work, encouraged by these advancements in improving the transferability of supervised models, we revisit supervision for continual representation learning. We argue that additional information (human annotations) should not hurt the quality of representations in continual learning, as suggested by Madaan et al. (2022). Motivated by the latest study on transferability of representations in self-supervised and supervised learning, we aim to improve transferability between tasks in continual learning. We are the first to show that supervised models can continually learn representations of higher quality than self-supervised models when trained with a simple MLP head (see Fig. 1).

The main contributions of this paper are as follows. (1) We empirically show that SL equipped with a simple MLP projector can learn higher-quality representations than SSL methods in continual finetuning scenarios. (2) We show that the use of the MLP projector can be coupled with several continual learning methods, further improving their performance. (3) We shed light on the reasons behind the strong performance of supervised learning with MLP projector: better transferability, lower forgetting, and increasing diversity of representations.

## 2 Related Work

**Self-supervised learning** In this work, we use BarlowTwins (Zbontar et al., 2021) (denoted as SSL) which considers an objective function measuring the cross-correlation matrix between the features and SimCLR (Chen et al., 2020) which uses contrastive learning based on noise-contrastive estimation. A number of studies (Bordes et al., 2023; Chen & He, 2020; Zbontar et al., 2021; Jing et al., 2022) show that an MLP projector between the encoder and the loss function is a crucial component to prevent the collapse of the representations and improve their transferability. **Transferable representations** Wang et al. (2021) found out that adding a projection network boosts the transferability of the supervised models' features as well. This was further explored in Sariyildiz et al. (2023) and it was shown that it is possible to build representations that are good for both the source and the downstream tasks. In this work, we revisit those findings in the context of models learned on a sequence of tasks.

**Supervised Continual Learning (SCL)** aims to create systems that can learn to solve novel tasks using new annotated data while retaining the ability to solve previously learned tasks (Parisi et al., 2019). **Unsupervised Continual Learning (UCL)** aims to improve the quality of learnt represetntations utilizing an ever-changing stream of unlabeled data. Recent works (Fini et al., 2022; Madaan et al., 2022; Gomez-Villa et al., 2021) apply SSL in the UCL setting and claim their superior results for continual representation learning.

## 3 Experimental Setup

**Datasets** We utilize four different datasets: CIFAR10 (Krizhevsky, 2009) (C10), CIFAR100 (Krizhevsky, 2009) (C100), SVHN (Netzer et al., 2011) and ImageNet100 (Tian et al., 2019) (IN100). We use $D/N$ to denote that dataset $D$ is split into $N$ tasks with an equal number of classes in each task without overlapping ones. We use $A \rightarrow B$ to denote task shift meaning that the model was trained on two tasks, the first one was dataset $A$ and the second one was dataset $B$.

**Methods** We use the following supervised methods: (1) SL - backbone with linear head with a cross-entropy loss function, (2) SL+MLP - SL with MLP projector added between the backbone and a linear head that is discarded at test-time, (3) t-ReX (Sariyildiz et al., 2023), and (4) SupCon (Khosla et al., 2020). For CL strategies we use LwF (Li & Hoiem, 2018), CaSSLe (Fini et al., 2022) and PRF (Gomez-Villa et al., 2021). We use ResNet-18 (He et al., 2016) as a feature extractor.

**Evaluation** We use k-NN classifier to evaluate the quality of representations following Fini et al. (2022); Madaan et al. (2022) and Nearest Mean Classifier (NMC) as in Rebuffi et al. (2017); Yu et al. (2020) to evaluate the stability of representations. We use CKA (Kornblith et al., 2019) to measure the similarity between representations of two models. Moreover, we use forgetting ($F$) and forward transfer ($FT$) commonly used in continual learning (Lopez-Paz & Ranzato, 2017). We also measure task exclusion difference $EXC$ (Hess et al., 2023) to evaluate the level of retention of task-specific features. We use subscripts to indicate the evaluation dataset, e.g. $Acc_{C10}$ means "accuracy on C10".

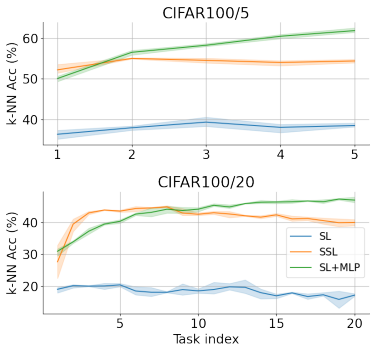


Figure 2: SL+MLP: (1) achieves strong performance after the initial task compared to SL which indicates that it produces representations that are transferable to the unseen tasks; (2) is the only method that is able to accumulate knowledge learned on a sequence of tasks. We report task-agnostic k-NN accuracy on the whole dataset (including unseen tasks).

| Method | CL strategy | C100/5 | C100/20 | IN100/5 |
|---|---|---|---|---|
| SUPERVISED CONTINUAL LEARNING | | | | |
| SL | Finetune | 38.5±0.4 | 17.2±0.3 | 35.3±1.3 |
| | LWF | 57.4±0.2 | 45.2±1.2 | 60.5±0.3 |
| | PFR | 57.7±0.4 | 44.4±1.3 | 58.7±0.2 |
| SL+MLP | Finetune | 61.9±0.5 | 47.1±0.7 | 62.4±0.4 |
| | LWF | 58.7±0.2 | 51.9±0.1 | 60.4±0.2 |
| | PFR | **63.6±0.2** | **54.5±0.2** | 65.2±0.1 |
| t-ReX | Finetune | 59.2±0.6 | 50.8±0.1 | 59.2±0.6 |
| | LwF | 58.3±0.4 | 50.4±0.1 | 58.6±1.0 |
| | PFR | 60.9±0.5 | 53.4±0.3 | 63.9±0.6 |
| SupCon | Finetune | 49.4±0.3 | 30.0±0.7 | 57.6±0.6 |
| | CaSSLe | 61.1±0.2 | 49.2±1.2 | **70.4±0.6** |
| | PFR | 57.0±0.2 | 51.2±0.8 | 68.0±0.7 |
| UNSUPERVISED CONTINUAL LEARNING | | | | |
| BarlowTwins | Finetune | 54.1±0.3 | 40.0±0.8 | 57.0±0.4 |
| | CaSSLe | **58.6±0.6** | 49.3±0.1 | **64.9±0.1** |
| | PFR | 57.2±0.2 | 46.0±0.7 | 61.1±0.2 |
| SimCLR | Finetune | 48.9±0.4 | 33.4±0.5 | 54.7±0.4 |
| | CaSSLe | 55.9±0.5 | 48.2±0.4 | 59.3±0.5 |
| | PFR | 53.8±0.3 | **49.4±0.1** | 57.7±0.2 |

Table 1: k-NN accuracy of the learnt representations. The best result in **bold** and second best underlined.

## 4 Experimental results

Fig. 2 presents our main results showing that supervised models can build stronger representations than self-supervised models under continual finetuning, contrary to previous beliefs (Madaan et al., 2022). We recognize that the key component to enhancing the performance of supervised models is an additional MLP projector during training and discarded afterward. We identify two factors contributing to superior results of SL+MLP. Firstly, we observe that the performance of supervised models after the initial task is largely improved by the addition of the MLP projector, resulting in accuracy close to SSL models. In order to achieve good task-agnostic accuracy on the whole dataset (seen and unseen classes), the model trained on a single task needs to perform well on unseen data. Therefore, we attribute the advantage of SL+MLP to the increased transferability of representations induced by MLP projector. Secondly, we notice that SL+MLP is the only method able to incrementally accumulate knowledge and consistently improve performance.

Tab. 1 presents extended results including multiple SL and SSL approaches in continual finetuning and paired with different CL methods. Firstly, we observe that all the supervised methods equipped with the projector (SL+MLP, t-ReX, and SupCon) significantly outperform simple SL. What is worth noting is the fact that all these methods were trained with different supervised losses: SL+MLP uses cross-entropy, t-ReX uses cosine softmax cross-entropy and SupCon uses supervised contrastive loss. Secondly, we observe that the positive effects of the MLP projector and CL strategy compound. As a result, the best models are those (1) trained in a supervised way (2) with the use of the MLP projector and (3) coupled with CL strategy based on temporal learnable projection, namely CaSSLe or PFR.

We investigate the **quality of representations** built by supervised and self-supervised training in Fig. 3. SL+MLP outperforms both SSL and SL in most of the experiments. In Tab. 2 we observe high **representation forgetting** for SL, significantly lower for SSL, and the lowest for SL equipped with MLP projector. It also shows the results of **task exclusion comparison** ($EXC$). SL achieves small positive $EXC$ meaning that it forgets most features specific to the initial task. SL+MLP achieves the highest $EXC$ which shows that it is able to successfully retain a large portion of task-specific features. Surprisingly, SSL exhibits negative $EXC$. It means that it is better to train SSL model from scratch on another task than to finetune the model pretrained on the task of interest. In Tab. 2 we report **CKA similarity** between the models trained on C10 and the rest of the models. We observe that usage of MLP head in SL increases CKA between the C10 model and other models. Moreover, the models pretrained on C10 and finetuned on another task have higher similarity to C10 models

| Training sequence | SL | | | | SSL | | | | SL+MLP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Acc_{C10}\uparrow$ | $F_{C10}\downarrow$ | $EXC_{C10}\uparrow$ | $\text{CKA}_{C10}\uparrow$ | $Acc_{C10}\uparrow$ | $F_{C10}\downarrow$ | $EXC_{C10}\uparrow$ | $\text{CKA}_{C10}\uparrow$ | $Acc_{C10}\uparrow$ | $F_{C10}\downarrow$ | $EXC_{C10}\uparrow$ | $\text{CKA}_{C10}\uparrow$ |
| C10 | 92.6±0.1 | - | - | - | 88.8±0.1 | - | - | - | **93.3±0.1** | - | - | - |
| C100 | 74.9±0.2 | - | - | 0.46±0.00 | 80.8±0.1 | - | - | **0.56±0.01** | **84.5±0.4** | - | - | 0.49±0.01 |
| C10→C100 | 76.1±0.1 | 16.6±0.2 | 1.2±0.3 | 0.50±0.00 | 79.1±0.2 | 9.7±0.3 | -1.8±0.2 | 0.52±0.01 | **88.8±0.2** | **4.5±0.3** | **4.3±0.6** | **0.57±0.00** |
| SVHN | 21.8±0.3 | - | - | 0.05±0.00 | **58.6±1.2** | - | - | **0.27±0.01** | 56.3±0.2 | - | - | 0.20±0.01 |
| C10→SVHN | 22.6±0.5 | 70.1±0.5 | 0.8±0.4 | 0.05±0.01 | 54.9±0.7 | 33.8±0.7 | -3.7±1.9 | **0.25±0.01** | **62.7±0.8** | **30.6±0.8** | **6.4±1.0** | **0.25±0.01** |

Table 2: We observe high representation forgetting for SL, significantly lower for SSL, and the lowest for SL+MLP. SL is able to preserve a small fraction of task-specific features while SL+MLP can retain much more, based on their $EXC$ scores. Surprisingly, SSL achieves negative $EXC$

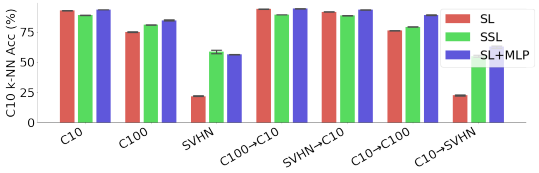


Figure 3: SL+MLP (blue) achieves better representations than SL (red) and SSL (green) in most sequences of tasks.

| Method | C10 | C100→C10 | | SVHN→C10 | |
|---|---|---|---|---|---|
| | $Acc_{C10}\uparrow$ | $Acc_{C10}\uparrow$ | $FT_{C10}\uparrow$ | $Acc_{C10}\uparrow$ | $FT_{C10}\uparrow$ |
| SL | 92.6±0.1 | 94.0±0.2 | **1.3±0.3** | 91.5±0.2 | -1.1±0.3 |
| SSL | 88.8±0.1 | 89.2±0.1 | 0.5±0.2 | 88.5±0.1 | -0.3±0.2 |
| SL+MLP | **93.3±0.1** | **94.3±0.1** | 1.0±0.0 | **93.2±0.2** | -0.1±0.1 |

Table 3: All methods benefit prom pretraining on C100 which is semantically close to C10. However, pretraining on semantically distant SVHN hinders the performance of SL.

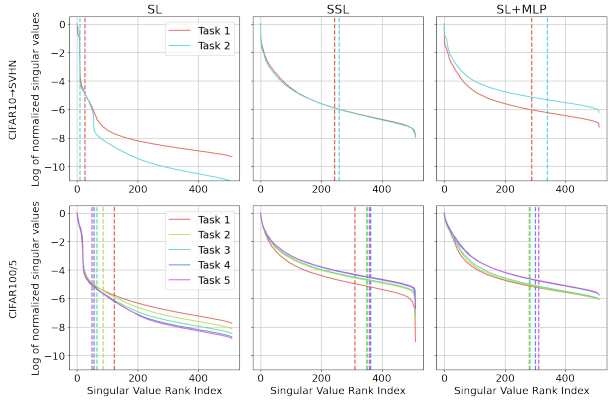


Figure 4: Representations learned with SL+MLP (right) exhibit desirable properties from the continual learning point of view: (1) they consist of a more diverse set of features (contrary to SL, left); (2) they improve feature diversity when learning new tasks consistently across all the presented settings. Vertical dashed lines denote 95% of the variance explained.

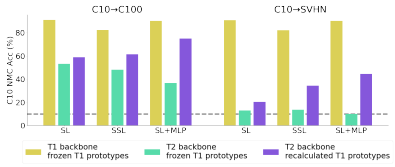


Figure 5: Task aware NMC accuracy on C10. After training on C10 (T1), both SL and SSL models achieve high NMC performance (yellow). After training the second task (T2), the nearest mean classification using old prototypes results in performance degradation (green). We calculate *an upper-bound* accuracy after training on the second task by recalculating the prototypes using old data and a new backbone (purple). Note that it is not possible in SCL as old data is inaccessible. Gray dotted line marks random guess performance.

than the models trained on another dataset from scratch, which is not necessarily the case for SL models. SSL models have the highest CKA scores, however, they usually underperform compared to SL+MLP suggesting that SSL produces similar but less discriminative features. We present the results of the **forward transfer** evaluation in Tab. 3. All the methods benefit from pretraining on CIFAR100 which is semantically close to CIFAR10. However, pretraining on semantically distant SVHN hinders the performance of SL but it hardly influences the performance of SSL and SL+MLP.

To gain further insight into the properties of continually trained representations, we analyze the **spectra of representations** following the procedure from Jing et al. (2022). We perform singular value decomposition of the covariance matrix of the representations $C = USV^T$, where $S = diag(\sigma^k)$ and $\sigma^k$ is $k$-th singular value of $C$. Fig. 4 presents how singular value spectra change after each task for different training methods and different sequences of tasks. Firstly, we observe that SL exhibits signs of neural collapse (Papyan et al., 2020) - a large fraction of variance is described by a few dimensions roughly equal to the number of classes in the training set. This is an undesirable property in continual representation learning as the representations should be more versatile and useful not only for current but also for past and future tasks. Adding MLP to SL prevents neural collapse and results in features' properties more similar to SSL. Secondly, we observe that for SL, the diversity of

the features decreases in subsequent tasks. SSL and SL+MLP are able to consistently improve the diversity of the representations suggesting its superiority in continual representation learning.

We define representations as **stable** when they do not drift in the representation space when the network is trained on a new task. The stability provides a different perspective when evaluating continually trained representations. It is irrelevant when the only objective is to continually learn representations, however, it is a desired property when we also want to solve downstream tasks continually (Yu et al., 2020). The results are presented in Fig. 5. Representations of all the methods are not stable in high distribution shift scenario C10→SVHN. However, in a low distribution shift scenario, C10→C100, SL exhibits high stability while SL+MLP underperforms in that regard. Note that performance degradation can be only partially attributed to forgetting of representations as the upper-bound performance is still high after training on the second task for most of the methods. These results suggest that there exists a trade-off between the stability and expressiveness of representations trained continually as methods that build stronger representations tend to have lower stability.

## Acknowledgments and Disclosure of Funding

Daniel Marczak is supported by National Centre of Science (NCN, Poland) Grant No. 2021/43/O/ST6/02482. This research was partially funded by National Science Centre, Poland, grant no 2020/39/B/ST6/01511 and grant no 2022/45/B/ST6/02817. Bartłomiej Twardowski acknowledges the grant RYC2021-032765-I.

## References

Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*, 2023.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Mohammad Reza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Linus Ericsson, H. Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Enrico Fini, Victor G. Turrisi da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal. Self-supervised models are continual learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Alex Gomez-Villa, Bartlomiej Twardowski, Lu Yu, Andrew D. Bagdanov, and Joost van de Weijer. Continually learning self-supervised representations with projected functional regularization. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altch'e, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, B. '. Pires, Z. Guo, M. G. Azar, Bilal Piot, K. Kavukcuoglu, R. Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *Neural Information Processing Systems (NeurIPS)*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

Timm Hess, Eli Verwimp, Gido M. van de Ven, and Tinne Tuytelaars. Knowledge accumulation in continually learned representations and the issue of feature forgetting. *arXiv preprint arXiv: 2304.00933*, 2023.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Neural Information Processing Systems (NeurIPS)*, 2020.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. *International Conference on Machine Learning (ICML)*, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 2009.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Divyam Madaan, Jaehong Yoon, Yuanchun Li, Yunxin Liu, and Sung Ju Hwang. Representational continuity for unsupervised continual learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 2019.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

Mert Bulent Sariyildiz, Yannis Kalantidis, Karteek Alahari, and Diane Larlus. No reason for no supervision: Improved generalization in supervised models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *European Conference On Computer Vision (ECCV)*, 2019.

Yizhou Wang, Shixiang Tang, Feng Zhu, Lei Bai, Rui Zhao, Donglian Qi, and Wanli Ouyang. Revisiting the transferability of supervised pretraining: an mlp perspective. *Computer Vision and Pattern Recognition (CVPR)*, 2021.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Computer Vision and Pattern Recognition (CVPR)*, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *International Conference on Machine Learning (ICML)*, 2021.

Nanxuan Zhao, Zhirong Wu, Rynson W. H. Lau, and Stephen Lin. What makes instance discrimination good for transfer learning? *International Conference on Learning Representations*, 2020.