Comparison of semi-supervised learning methods for High Content Screening quality control

Anonymous ECCV submission

Paper ID 3

Abstract. Progress in automated microscopy and quantitative image analysis has promoted high-content screening (HCS) as an efficient drug discovery and research tool. While HCS offers to quantify complex cellular phenotypes from images at high throughpout, this process can be obstructed by image aberrations such as out-of-focus image blur, fluorophore saturation, debris, a high level of noise, unexpected autofluorescence or empty images. While this issue has received moderate attention in the literature, overlooking these artefacts can seriously hamper downstream image processing tasks and hinder detection of subtle phenotypes. It is therefore of primary concern, and a prerequisite, to use HCS. In this work, we evaluate deep learning options that do not require extensive image annotations to provide a straightforward and easy to use semi-supervised learning solution to this issue. Concretely, we compared the efficacy of recent self-supervised and transfer learning approaches to provide a base encoder to a high throughpout artefact image detector. The results of this study suggest that transfer learning methods should be preferred for this task as they not only performed best here but present the advantage of not requiring sensitive hyperparameter settings nor extensive additional training.

Keywords: Cell-based assays · Image analysis · Deep learning · Self-supervised learning

1 Introduction

Image analysis solutions are heavily used in microscopy. They enable the extraction of quantitative information from cells, tissues and organisms. These methods and tools have proven to be especially useful for high-content screening (HCS), an automated approach that produces a large amount of microscopy image data, to study various mechanisms and identify genetic and chemical modulators in drug discovery and research [19]. However, the success of an HCS screen is often related to the dataset quality obtained at end. In practice, abnormalities in image quality are numerous and can lead to imprecise results at best, and erroneous results or false conclusions at worst. Common abnormalities include noise, out-of-focus, presence of debris, blur or image saturation. Furthermore, in some cases, it can also be convenient to exclude images full of dead or floating cells. More importantly, in HCS, manual inspection of all images in a dataset is

intractable, as one such screen typically encompasses hundreds of thousands of images.

Quality control (QC) methods have been investigated for this purpose. Interesting software such as CellProfiler [6] allows end-to-end analysis pipeline with an integrated QC modules. Although powerful, the image quality measures are mainly handcrafted with different computed metrics as described in [2] and therefore hard to generalize. More recently, Yang et al proposed a method to assess microscope image focus using deep learning [21]. However, this approach is restricted to a specific type of aberration and does not generalize well to other kinds of artefacts. Besides, learning all types of aberrations from scratch in a supervised manner is hardly tractable, given the diversity of both normal and abnormal image types. It would require systematic annotation of all types of aberrations on each new high-throughput assay, and thus would be utterly time-consuming and hardly feasible in practice. For this task, we thus typically seek a semi-supervised solution that would require annotation of a limited amount of data per assay.

Transfer learning typically offers such a solution that relies on little supervision [13]. A network pretrained on a large annotated image set can be reused directly or fine-tuned with a limited set of annotated images to solve a specific task in another domain. Furthermore, recent breakthroughs in self-supervised learning (SSL), which aim to learn representations without any labels data call for new methods [12, 9, 23, 1, 4, 20, 5, 7]. For instance, such a framework was successfully used by Perakis et al [17] to learn single-cell representations for classification of treatments into mechanisms of action. It was shown that SSL performed better than the more established transfer learning (TL) in several applications. However, it is not a strict rule and not systematically the case as assessed by a recent survey [22]. It is still unclear which approach works better on what type of data and tasks.

In this work, we propose to address this question in the context of HCS quality control. To this end we performed a comparative study of a range of SSL and TL approaches to detect abnormal single-cell images in a high-content screening dataset with a low amount of annotated assay specific image data. The paper is organized as follows. In Section 2, we briefly describe the various methods we use for transfer and self-supervised representation learning. In Section 3, we then detail the setup of this comparative study. We then provide experimental results in Section 4, and concluding remarks in Section 5.

2 Related work

We seek a method that would provide a robust base encoder to a quality control downstream task where a low amount of annotated data is available. We thought of several options that could be grouped in two categories: transfer learning and self-supervised learning methods.

2.1 Transfer learning

Training a deep learning model efficiently necessitates a significant amount of data. In the case of supervised training, it is required that data be annotated with class labels. Transfer learning has become popular to circumvent this issue. It consists in pretraining a network on a large set of annotated images in a given domain, typically a domain where image could be annotated. A variety of tasks in various other domains can then be addressed with decent performance simply by reusing the pretrained network as is or by fine tuning its training on a small available dataset on a specific task in the domain of interest.

In this work, we included three popular networks pretrained with ImageNet for transfer learning. First we used VGG16, a model introduced in 2014 that made a significant improvement over the early AlexNet introduced in 2012, by widening the size of convolutional layer kernels [18]. We also use ResNet18, a network introduced in 2016 that implements residual connections to make possible the stable training of deeper networks [11]. Finally we use ConvNext, one of the most recent convolutional networks introduced in 2022 that competes favorably with most models, including vision transformer, while maintaining the simplicity and efficiency of ConvNets [14].

2.2 Self-supervised learning

In recent years, self-supervised representation learning has gained popularity thanks to its ability to avoid the need for human annotations. It has provided ways to learn useful and robust representation without labeling any data. Most of these approaches rely on a common and simple principle. Two or more random transformations are applied to the same images to produce a set of images containing different views of the same information content. These images are then passed through an encoder that is trained to somewhat encourage learning of a close and invariant representation through the optimization of a given loss function. The loss function varies depending on the method, but once a self-supervised representation is learned, it can be used to solve downstream tasks that may require little to no annotated data. Various kind of SSL mechanisms have been developed, but a wide range of approaches can be summarized in three classes of methods our study encompasses here, namely contrastive, non contrastive and clustering-based methods:

1. **Contrastive learning methods** aim to group similar samples closer and diverse samples farther from one another. Although powerful, such methods still need to find some negative examples via a memory bank or to use a large batch size for end to end learning [12].
2. **Non-contrastive learning methods** use only positive sample pairs compared to contrastive methods. These approaches proved to learn good representations without the need for a large batch size or memory bank [9, 23, 1].

3. **Clustering-based learning methods** ensure that similar samples cluster together but use a clustering algorithm instead of similarity metrics to better generalize by avoiding direct comparison [4, 20, 5].

In this work, we used a list of methods from the three framework listed above, namely SimCLR [7] for contrastive learning (based on similarity maximisation objective), Barlow Twins [23] and VICReg [1] for non-contrastive techniques (based on redundancy reduction objective) and DeepCluster [4] and SwAV [5] for clustering-based methods.

3 Method

We performed a comparative study that aimed at identifying which of the previously described approaches could be best suited to provide a base encoder, in order to build a classifier for abnormal images from a small annotated image set. In this section we describe the data, the way we perform training for the encoders and the downstream tasks we designed to evaluate and compare them.

3.1 Data

We used the BBBC022 image set, available from the Broad Bioimage Benchmark Collection to conduct our experiments [10, 16]. To obtain images at a single-cell level, we cropped a fixed 128×128 pixel square around the center of each nucleus, resulting in a total of $2,122,341 (\approx 2.1M)$ images [15] . Most of these images were used to train the base encoder when needed (i.e. for SSL methods). Separately, we manually annotated 240 abnormal and 240 normal images and split them into training (350) and test (130) sets for the downstream tasks. Some annotated images are displayed in Figure 1. Furthermore, we also used 200 annotated images from the BBBC021 image set, available from the Broad Bioimage Benchmark Collection to test the generalization of our approach [3, 16].

3.2 Encoder training

For all TL methods, we used a model pre-trained on ImageNet as encoder. For SSL methods, we used two networks - first a ResNet18 as encoder and a fully connected layers (FC layers) as projector. Once the encoder was trained, the projector network was discarded and only the ResNet18 network was used for downstream tasks. The decoder takes an image as input and outputs a 512 dimension vector. We forward pass the batch of images after producing two different views of them using augmentations. The following augmentations were randomly performed: 90 degree rotation, flip, transpose, shift and scale. We carefully chose these augmentations so as to keep the trained features relevant to our downstream classification tasks. The encoder was trained for 5 epochs (about 10 million images) for all the SSL methods with a batch size of 128,

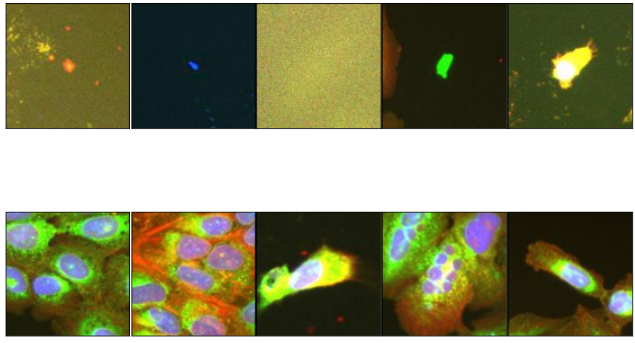

Fig. 1. BBBC022 dataset: the first row displays a few abnormal images, the second row shows a few regular images of cells. U2OS cells with Hoechst 33342 staining for nuclei (blue), WGA + phalloidin staining for actin filaments (Red) and MitoTracker staining for mitochondria (green).

using the SGD optimizer with an initial learning rate of 0.001 and a momentum of 0.9. We also used a warm-up cosine scheduler with warm-up epochs set to 1. The projector network was made of 2048 dimension FC layers and a temperature of 0.07 for the unsupervised loss. For DeepCluster, we set 500 prototypes and 10 K-means iterations. For SwAV we set the number of prototypes to 500 and choose a queue size of 2560. Other projector network parameters were the same as those used in the original papers. We trained models and ran experiments using one Tesla P100 GPU with 16GB vram. All experiments involving SSL methods were done using the solo-learn library [8].

3.3 Downstream Classification tasks

After training an encoder with each previously described TL or SSL method, we used them to train and test the three following downstream classification tasks, only with the small annotated dataset previously described:

1. **K-nn on a freezed encoder output.** We first aimed to evaluate a simple classification setting that did not necessitate any additional training. To this end, we performed a K-Nearest Neighbour (KNN) classification (here we chose k=5) on the 512 feature vectors output of the encoder.
2. **Linear classifier trained on a freezed encoder output.** We then evaluated the supervised training of a single dense layer with 2 output classes on top of the pre-trained encoder. In this setting, the weights of the pre-trained encoder were frozen. We trained this layer for 150 epochs and with a batch size of 32. The optimizer was SGD with a learning rate set to 0.001 and momentum to 0.9. We also used a step scheduler with gamma value set to 0.1 at 40, 70 and 120 epochs.
3. **Linear classifier with fine tuned training of the encoder.** We then used a dense linear layer with 2 output classes as in the previous settings.

However, this time we did not freeze the encoder network and allowed it to pursue training. We trained the models for 50 epochs. The learning rate was 0.001 with a momentum of 0.9 with a SGD optimizer. We also used a step scheduler with a gamma value of 0.1 at 25 on 40 epochs.

3.4 Evaluation Criteria

We used Accuracy, F1-Score and the Area Under Curve (AUC) score to assess the classification results. All displayed values are weighted average for the 2 classes. The most important metric is the F1 score because it takes both false positives and false negatives into account. Thus, the higher the F1 score, the better the result. All values mentioned are in percentages.

4 Results

4.1 Evaluations on downstream tasks

The results for the Linear Layer classifier and KNN are displayed in Table 1. Among the self-supervised method, we can observe that DeepCluster performs best in both settings and reaches a maximum of 94.57% accuracy. SimCLR also performs best with KNN while VICReg performs poorly, dropping to 76.30% accuracy. However, none of the SSL methods outperforms the three TL encoders with ConvNext culminating at 98.47% accuracy with a Linear Layer classifier.

Furthermore, the results obtained with fine tune trainings of all the encoder weights are displayed in the first 3 columns of Table 2. We can see that with 350 training images (the full annotated training image set), the best results were again obtained with the three TL methods. However, SSL method performed almost as well in this setting with simCLR reaching the best results among the SSL methods with 98.44% accuracy.

4.2 Effect of a decreasing amount of annotated data

We also performed an ablation study where the number of training images was gradually decreased. We performed training of the third task with 350, 100, 50, 25, and finally just 10 images. The purpose for decreasing the amount of training images was to evaluate how much supervision the network needs to perform properly.

The results are displayed in Table 2. As the number of training images decrease, VICReg, DeepCluster and SwAV display a drop in performance. With only 25 images, Barlow Twins still produces fairly good results with 94.53% accuracy. With just 10 images, the best result among the SSL methods is simCLR with 84.89% accuracy. Overall, semi-supervised training can yield good results even with a few images. However, here again, none of the SSL methods outperforms the transfer learning baselines.

Table 1. Classification with KNN or a single Linear Layer with a frozen encoder using a 350-image training set. 130 images were used for test.

	KNN			Linear Layer		
Method	Acc	F1	AUC	Acc	F1	AUC
VGG16	96.09	95.94	95.38	98.24	98.12	98.09
ResNet18	**97.66**	**97.51**	**97.63**	98.44	98.21	**98.26**
ConvNext	97.65	97.42	97.56	**98.47**	**98.24**	98.17
SimCLR	90.62	89.63	90.30	91.40	90.56	91.20
Barlow Twins	89.84	88.02	88.24	87.28	86.42	87.35
VICReg	76.30	75.28	75.13	87.76	87.36	87.40
DeepClusterV2	90.62	89.50	89.81	94.57	94.02	94.14
SwAV	89.84	88.52	89.46	94.53	94.00	94.08

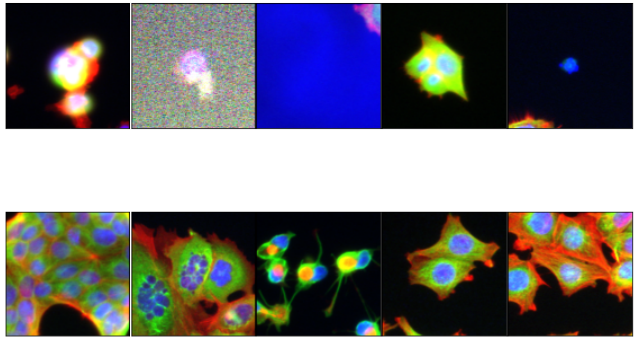

Fig. 2. BBBC021 dataset: the first row displays a few abnormal images, the second row displays a few regular cell images. Fixed human MCF7 cells labeled for DNA (blue), actin (red) and B-tubulin (green)

4.3 Effect of a domain shift

To evaluate how these encoders pretrained on BBBC022 or ImageNet could generalize to a different dataset, we tested them on data taken from BBBC021. For this purpose, we considered our best model, the Linear Layer approach with fine tune training of all the encoder weights on 350 images from the BBBC022 dataset. We then tested it on unseen data taken from BBBC021. We annotated 100 normal and 100 abnormal images from this last dataset for this purpose. Some sample images are displayed in Figure 2. We made sure to include diverse images in order to thoroughly check the robustness of our trained models.

The results are displayed in Table 3. Among SSL methods, SimCLR and DeepCluster performed best with respectively 73.66% and 72.32% accuracy. These results show that some self-supervised learning methods such as simCLR or DeepCluster trained on a large dataset produce features that could generalize

a quality control task to an unseen dataset to some extent. However, in accordance with what was observed in previous sections on BBBC022, none of these approaches outperformed the results obtained with the TL encoders.

Table 2. Effect of a decreasing amount of training images on a Linear Layer classifier with a non-frozen encoder. 130 images were used for test.

	Number of Training Images														
	350			100			50			25			10		
Method	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
VGG16	99.08	98.92	99.01	99.12	98.83	98.89	**99.22**	**99.18**	**99.24**	**98.44**	**98.19**	**98.26**	96.09	95.50	95.81
ResNet18	99.19	99.11	99.02	99.08	98.96	99.07	98.54	98.48	98.47	97.66	97.81	97.62	93.75	93.83	93.54
ConvNext	**99.39**	**99.21**	**99.25**	**99.28**	**99.17**	**99.24**	97.66	97.27	97.53	97.62	97.49	98.02	**96.87**	**96.68**	**96.55**
SimCLR	98.44	98.21	98.34	91.93	92.52	93.00	92.97	93.13	93.15	92.97	93.17	92.80	84.89	84.61	84.32
Barlow Twins	98.44	98.14	98.12	95.05	95.08	94.95	94.53	94.21	93.86	94.53	94.12	93.65	75.00	80.39	83.14
VICReg	98.24	98.00	97.88	89.32	89.03	89.20	71.01	74.33	74.15	72.13	77.37	80.22	66.40	72.12	72.00
DeepClusterV2	96.87	96.18	96.25	81.77	82.47	83.34	83.59	83.00	83.20	86.98	86.53	87.09	83.13	83.31	82.79
SwAV	94.53	94.00	94.10	83.59	83.55	83.70	82.56	82.21	83.09	87.50	87.12	87.25	82.87	82.62	83.39

Table 3. Out of Domain Test. Linear Layer classifier with a non-frozen encoder trained on 350 images from the BBBC022 dataset and tested on the 200 images of the BBBC021 dataset.

	Linear Layer		
Method	Acc	F1	AUC
VGG 16	96.43	97.51	98.02
ResNet18	91.52	93.21	92.87
ConvNext	**98.66**	**98.53**	**98.91**
SimCLR	73.66	78.47	79.00
Barlow Twins	54.91	62.02	58.76
VICReg	37.95	40.00	39.21
DeepClusterV2	72.32	75.99	76.12
SwAV	56.25	57.50	57.30

5 Conclusion

In this work, we conducted a thorough investigation to evaluate transfer and self-supervised representation learning on a large dataset in order to perform a downstream HCS quality control task. The quantitative results we obtained suggest that TL approaches perform better than SSL for this task. Importantly, all SSL methods come with the need to choose crucial hyperparameters that

will have significant impact on the learned representation. Among these hyperparameters are the choice of transformations that will define feature invariance in the obtained representation. Furthermore, SSL methods require an additional training on a large set of unannotated images. In contrast, an ImageNet pretrained encoder combined with a KNN downstream can be used out of the box and does not require any training or hyperparameter setting. If training can be performed, then unfreezing the encoder weights and fine tuning the training with a low amount of annotated data will slightly increase the performances, with TL still being a better option than SSL. Altogether this suggests that for the task of identifying abnormal versus normal image, transfer learning should be the preferred choice.

Two reasons could be hypothesized to explain our findings. First, one could argue that our choice of transformations for the SSL approaches may not be the best option to create an optimal representation for our downstream quality control tasks. However, the choices we made were reasonable and relevant, and anyone seeking to solve a task using SSL would face the same issue: choosing hyperparameters and performing an additional training. Importantly, the debate on hyperparameter settings would be sound if transfer learning did not perform so well. Here we show that it is not only performing better than all SSL approaches, but it reaches almost perfect results in several setups, suggesting that even a better choice of SSL augmentations would not necessarily be worth finding. Secondly, this high performance obtained with transfer learning may be related to the specificity of the downstream task. Indeed, the experiments performed in the papers presenting these SSL approaches are often based on ImageNet classification which contains homogeneous semantic classes and therefore represents a different objective than the one presented in this work. Abnormal images do represent a very variable class with, for instance, out-of-focus image of cells being very different than an image containing debris. In this case, the low level features retrieved from the natural images of ImageNet may simply be sufficient and more efficient than higher semantic structure SSL representation typically provides. Although we focused on high-content screening here, we hope our findings will benefit quality control in other imaging modalities.

References

1. Bardes, A., Ponce, J., Lecun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations (2022)
2. Bray, M.A., Carpenter, A.E.: Quality control for high-throughput imaging experiments using machine learning in cellprofiler. Methods in molecular biology **1683**, 89–112 (2018)
3. Caie, P.D., Walls, R.E., Ingleston-Orme, A., Daya, S., Houslay, T., Eagle, R., Roberts, M.E., Carragher, N.O.: High-content phenotypic profiling of drug response signatures across distinct cancer cellsphenotypic profiling across cancer cell types. Molecular cancer therapeutics **9**(6), 1913–1926 (2010)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European conference on computer vision (ECCV). pp. 132–149 (2018)
5. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
6. Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P., Sabatini, D.M.: Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology **7**, R100 – R100 (2006)
7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
8. da Costa, V.G.T., Fini, E., Nabi, M., Sebe, N., Ricci, E.: solo-learn: A library of self-supervised methods for visual representation learning. J. Mach. Learn. Res. **23**, 56–1 (2022)
9. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
10. Gustafsdottir, S.M., Ljosa, V., Sokolnicki, K.L., Anthony Wilson, J., Walpita, D., Kemp, M.M., Petri Seiler, K., Carrel, H.A., Golub, T.R., Schreiber, S.L., et al.: Multiplex cytological profiling assay to measure diverse cellular states. PloS one **8**(12), e80999 (2013)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)
13. Kensert, A., Harrison, P.J., Spjuth, O.: Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. SLAS Discov **24**(4), 466–475 (Apr 2019)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
15. Ljosa, V., Caie, P.D., ter Horst, R., Sokolnicki, K.L., Jenkins, E.L., Daya, S., Roberts, M.E., Jones, T.R., Singh, S., Genovesio, A., Clemons, P.A., Carragher,

N.O., Carpenter, A.E.: Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. Journal of Biomolecular Screening **18**, 1321 – 1329 (2013)
16. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature Methods **9**, 637–637 (2012)
17. Perakis, A., Gorji, A., Jain, S., Chaitanya, K., Rizza, S., Konukoglu, E.: Contrastive learning of single-cell phenotypic representations for treatment classification. International Workshop on Machine Learning in Medical Imaging (2021)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv (2014)
19. Singh, S., Carpenter, A.E., Genovesio, A.: Increasing the content of high-content screening. Journal of Biomolecular Screening **19**, 640 – 650 (2014)
20. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Proesmans, M., Van Gool, L.: Scan: Learning to classify images without labels. In: European conference on computer vision. pp. 268–285. Springer (2020)
21. Yang, S.J., Berndl, M., Ando, D.M., Barch, M., Narayanaswamy, A., Christiansen, E.M., Hoyer, S., Roat, C., Hung, J., Rueden, C.T., Shankar, A., Finkbeiner, S., Nelson, P.: Assessing microscope image focus quality with deep learning. BMC Bioinformatics **19** (2018)
22. Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S., Xie, P.: Transfer learning or self-supervised learning? a tale of two pretraining paradigms. ArXiv (2020)
23. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320. PMLR (2021)