

Self-Supervised Learning Strategies for Representation Extraction From Unlabeled Image Datasets

Dr. Chimi Yangzom

Lecturer, Institute of Intelligent Image Analysis, Bhutan

Abstract

Large image collections can often be acquired more easily than reliable human-generated labels. Conventional supervised learning depends on annotated examples, creating substantial labeling costs when datasets contain thousands or millions of images. Self-supervised learning addresses this limitation by deriving supervisory signals directly from the structure of unlabeled data. Visual representations can therefore be learned through reconstruction, transformation prediction, contrastive objectives, clustering, or self-distillation before being transferred to downstream recognition tasks. This hypothetical methodological study compares four self-supervised representation-learning strategies: image reconstruction, rotation prediction, contrastive representation learning, and bootstrap-based non-contrastive learning. A fictional corpus containing 24,000 unlabeled images is used for conceptual representation pretraining. A separate hypothetical downstream dataset containing 4,000 labeled images is employed only after pretraining to evaluate representation quality through linear classification, nearest-neighbor recognition, and limited-label fine-tuning. The illustrative analysis assigns the strongest frozen-feature performance to the bootstrap-based and contrastive strategies, while reconstruction produces the weakest semantic transfer results.

A second comparison evaluates label efficiency using one, five, ten, and twenty-five percent of the available downstream labels. Under the assigned hypothetical values, self-supervised initialization improves performance particularly strongly when labeled information is limited. These numerical outcomes are constructed examples and were not generated through actual neural-network training. The study concludes that self-supervised learning offers a valuable methodological approach for extracting reusable representations from unlabeled image collections. However, performance depends substantially on augmentation design, pretext-task choice, architecture, optimization stability, representation collapse prevention, and the similarity between pretraining and downstream data. Empirical validation using real image collections and repeated controlled experiments is required before performance superiority can be established.

Keywords: Self-Supervised Learning, Representation Learning, Unlabeled Images, Contrastive Learning, Computer Vision, Feature Extraction, SimCLR, BYOL, Image Representation, Deep Learning.

1. Introduction

Deep learning has produced substantial advances in computer vision, but conventional supervised models commonly require large collections of accurately labeled images. Creating such annotations can be costly because human reviewers must inspect individual images and assign appropriate categories, boundaries, or attributes. In many practical environments, unlabeled images accumulate far more rapidly than annotated examples. Representation learning from unlabeled data therefore offers an important approach for reducing dependence on manual supervision while preserving the ability to transfer learned features to later visual tasks.

Earlier unsupervised representation-learning methods frequently used autoencoders, clustering, dimensionality reduction, or reconstruction objectives. Hinton and Salakhutdinov demonstrated that neural networks could learn compressed representations of high-dimensional data, while Vincent et al. developed denoising autoencoders that reconstructed clean inputs from corrupted observations. Such methods demonstrated that useful structure can be learned without semantic labels. However, reconstruction objectives may emphasize low-level visual information such as texture and color rather than features most useful for object-level recognition.

Self-supervised learning developed by constructing surrogate or pretext tasks in which labels are generated automatically from the input data. Dosovitskiy et al. learned visual features by distinguishing transformed image patches, while Doersch et al. trained networks to predict spatial relationships between image regions. Pathak et al. used missing-region reconstruction, and Noroozi and Favaro introduced jigsaw-puzzle prediction. These approaches demonstrated that carefully designed pretext tasks can force neural networks to learn transferable structure from otherwise unlabeled images.

More recent approaches increasingly rely on invariance between transformed views of the same image. Contrastive methods such as Momentum Contrast and SimCLR encourage representations of related views to become similar while distinguishing different images. Other strategies, including BYOL and SimSiam, demonstrate that useful representations can also be learned without explicit negative examples when architectural or optimization mechanisms prevent representation collapse.

The present hypothetical study compares several major self-supervised strategies under a common methodological framework. Reconstruction, transformation prediction, contrastive learning, and bootstrap-based representation learning are evaluated through frozen-feature classification and limited-label fine-tuning. The objective is not to reproduce benchmark results from established datasets but to illustrate how the quality and label efficiency of representations extracted from unlabeled images can be assessed systematically.

2. Objectives of the Study

The primary objective of this study is to examine alternative self-supervised strategies for extracting transferable visual representations from unlabeled image datasets. Particular attention is given to representation quality, downstream transfer, label efficiency, augmentation design, and methodological challenges that influence the reliability of self-supervised visual learning.

2.1 Examination of Self-Supervised Representation Strategies

The first objective is to compare conceptually distinct mechanisms for generating learning signals without manually supplied class labels. Reconstruction-based learning attempts to recover image information, rotation prediction requires recognition of applied geometric transformations, contrastive

learning compares augmented views, and bootstrap-based learning aligns representations generated by related network branches. Each strategy imposes different assumptions concerning which visual characteristics should remain encoded within the learned feature space.

The study also examines whether a pretraining task encourages semantic representations or primarily preserves low-level image properties. A network may become highly effective at reconstructing pixels without learning features that separate meaningful object categories. Downstream evaluation is therefore essential because the training objective itself does not determine representation usefulness. A transferable representation should support new recognition tasks using relatively little labeled information rather than simply minimizing its original self-supervised loss.

2.2 Evaluation of Transfer and Label Efficiency

The second objective is to assess representation quality through downstream recognition after self-supervised pretraining. A frozen encoder is evaluated through linear classification and nearest-neighbor recognition so that differences can be attributed primarily to learned feature quality. Full or partial fine-tuning is considered separately because adaptation can compensate for weaker initial representations. This distinction provides a more informative comparison than evaluating only a fully optimized downstream network.

The study further examines label efficiency by fine-tuning the pretrained representations with progressively larger fractions of labeled downstream data. Self-supervised learning is especially valuable when annotation availability is limited. Consequently, performance with one or five percent of downstream labels provides an important test of representation usefulness. The analysis illustrates how pretraining strategies can be compared according to their ability to reduce the amount of supervised information required for acceptable recognition performance.

2.3 Identification of Self-Supervised Learning Challenges

The third objective is to identify methodological problems that influence self-supervised visual representation learning. Important concerns include inappropriate data augmentation, representation collapse, computational cost, domain mismatch, shortcut learning, and sensitivity to architecture or batch configuration. Self-supervised systems can learn features that satisfy the pretraining objective without capturing the semantic information required by downstream applications. Careful experimental design is therefore necessary.

The objective also considers evaluation fairness. Competing self-supervised approaches should use the same backbone architecture, pretraining images, downstream split, and evaluation protocol wherever possible. Larger networks or longer training schedules can improve performance independently of the learning objective. Comparative conclusions should therefore distinguish methodological improvements from differences in computational resources, data augmentation, model capacity, and optimization effort.

3. Review of Literature

3.1 Early Unsupervised and Pretext-Based Visual Learning

Hinton and Salakhutdinov demonstrated neural dimensionality reduction using deep autoencoders, while Vincent et al. developed denoising autoencoders that learned representations through reconstruction of corrupted inputs. Coates et al. later showed that relatively simple unsupervised feature-learning

procedures could perform competitively when appropriate preprocessing and feature extraction were used. These approaches helped establish the principle that useful representations can emerge without manually annotated semantic categories.

Visual self-supervision later shifted toward tasks created automatically from image structure. Dosovitskiy et al. proposed exemplar-based convolutional learning, and Doersch et al. used relative patch position as a supervisory signal. Pathak et al. developed context encoders for predicting missing image regions. Noroozi and Favaro introduced jigsaw puzzles, while Zhang et al. demonstrated that image colorization could function as a representation-learning task. These methods showed that pretext-task design strongly influences the features acquired by a network.

3.2 Contrastive and Invariance-Based Learning

Contrastive representation learning aims to bring related views closer in embedding space while separating unrelated observations. He et al. introduced Momentum Contrast, using a queue and momentum encoder to maintain a large and relatively consistent dictionary of representations. Chen et al. developed SimCLR and demonstrated the importance of augmentation composition, projection heads, batch size, and training duration. These approaches substantially strengthened the performance of self-supervised representations on standard downstream recognition protocols.

Tian et al. examined what constitutes an effective pair of views for contrastive learning and emphasized that augmentations should remove nuisance information while retaining task-relevant content. Contrastive learning therefore depends on more than the loss function itself. If two views remove semantically important information, the model may learn undesirable invariances. Conversely, weak augmentations can make the self-supervised task too simple and encourage representations based on superficial similarities rather than robust visual structure.

3.3 Non-Contrastive and Clustering-Based Strategies

Grill et al. introduced BYOL, in which an online network predicts the representation produced by a slowly updated target network without requiring explicit negative pairs. Caron et al. developed SwAV, combining clustering and consistency across multiple augmented views. Chen and He later proposed SimSiam, showing that a comparatively simple Siamese architecture with stop-gradient operations could learn useful representations without negatives or a momentum encoder. These developments broadened the design space beyond conventional contrastive objectives.

Zbontar et al. introduced Barlow Twins, using cross-correlation and redundancy reduction to encourage invariant but nonredundant representations. Caron et al. later demonstrated DINO, a self-distillation approach that produced strong representations with Vision Transformers and exhibited emergent semantic properties. Collectively, these methods demonstrate that self-supervised representation learning can be formulated through several mechanisms, including contrast, clustering, prediction, redundancy reduction, and teacher–student consistency.

4. Research Methodology

4.1 Research Design and Self-Supervised Models

The present study adopts a hypothetical comparative methodological design involving four self-supervised pretraining strategies. The first uses an autoencoder-style reconstruction objective. The second predicts image rotation among four predefined orientations. The third uses a SimCLR-style

contrastive objective applied to two stochastic augmentations of each image. The fourth uses a BYOL-style bootstrap objective with online and target encoders. All configurations conceptually use the same convolutional backbone and identical unlabeled image corpus.

Pretraining is conducted without access to downstream class labels. The learned encoder from each configuration is subsequently transferred to a separate labeled evaluation dataset. The first evaluation freezes encoder parameters and learns only a linear classification head. A nearest-neighbor evaluation is also conceptually performed directly in representation space. A second evaluation fine-tunes the pretrained model with restricted quantities of labeled data to examine label efficiency.

4.2 Hypothetical Image Dataset and Augmentation Procedure

The fictional pretraining corpus contains 24,000 unlabeled RGB images representing diverse objects, scenes, textures, and natural environments. Class labels are assumed unavailable during self-supervised training. A separate downstream dataset contains 4,000 labeled images across ten fictional recognition categories, divided into 2,000 training images, 1,000 validation images, and 1,000 test images. None of these records corresponds to an existing public or proprietary image dataset.

Contrastive and bootstrap pretraining conceptually use random resized crops, horizontal reflection, moderate color transformations, and limited grayscale conversion. Rotation prediction receives geometrically transformed versions of the original image, while reconstruction receives corrupted inputs from which original visual information must be recovered. Augmentation parameters are fitted only during pretraining and are applied independently to each view. Extreme transformations that could change the semantic identity of the image are deliberately excluded from the conceptual design.

4.3 Evaluation Metrics and Analytical Procedure

Representation quality is assessed using linear-probe classification accuracy, five-nearest-neighbor accuracy, and macro F1 score on the hypothetical downstream test set. The encoder remains frozen during linear and nearest-neighbor evaluation, allowing the metrics to reflect information already encoded during self-supervised pretraining. A randomly initialized encoder is included as a non-pretrained reference so that improvements associated with representation learning can be distinguished from downstream classifier capacity alone.

Label efficiency is evaluated by fine-tuning models using one, five, ten, and twenty-five percent of the 2,000 labeled downstream training images. Identical label subsets are assumed across pretraining strategies. Classification accuracy is reported after hypothetical fine-tuning. All numerical values presented below are deliberately assigned examples. No actual image encoder, self-supervised optimization procedure, or downstream classifier was executed to produce these results.

5. Analysis

5.1 Comparative Quality of Learned Representations

The first hypothetical analysis compares frozen representations across five initialization conditions. The randomly initialized reference produces the weakest linear and nearest-neighbor performance. Reconstruction-based pretraining improves the representation, followed by rotation prediction. Contrastive learning and bootstrap-based learning record the strongest illustrative downstream scores, with the bootstrap model reaching a linear-probe accuracy of 83.7% and macro F1 of 83.1%.

These assigned differences illustrate how downstream semantic evaluation can distinguish representations that may have achieved equally successful pretext-task optimization. Reconstruction may preserve substantial pixel-level information without developing the same category-level separation as invariance-based strategies. However, the ranking is entirely hypothetical. Real outcomes depend on image diversity, network architecture, optimization schedule, augmentation design, and downstream domain similarity, and therefore require repeated empirical experimentation.

Table 1: Hypothetical Downstream Performance of Self-Supervised Representations

Representation Strategy	Linear-Probe Accuracy (%)	5-NN Accuracy (%)	Macro F1 (%)
Random Initialization	52.8	48.6	51.9
Reconstruction Pretraining	68.4	64.7	67.8
Rotation Prediction	74.6	70.9	74.0
Contrastive Learning	82.1	78.8	81.5
Bootstrap Self-Supervision	83.7	80.2	83.1

5.2 Label Efficiency of Self-Supervised Initialization

The second hypothetical analysis evaluates fine-tuned recognition accuracy as the quantity of labeled downstream training data increases. The random initialization baseline records only 41.8% accuracy with one percent of labels, whereas contrastive and bootstrap initialization reach 63.8% and 65.4%, respectively. Differences narrow when twenty-five percent of labels are available, illustrating conceptually that self-supervised pretraining may provide its greatest practical value under limited supervision.

The assigned pattern demonstrates why low-label evaluation is important in self-supervised research. A representation that provides only a small advantage after extensive supervised fine-tuning can still be valuable if it substantially improves performance when annotation is scarce. Nevertheless, label efficiency should be assessed across repeated subsets because performance can vary depending on which labeled examples are selected. The table below is an illustrative example rather than a measured learning curve.

Table 2: Hypothetical Fine-Tuning Accuracy Under Limited Labeled Data

Pretraining Strategy	1% Labels	5% Labels	10% Labels	25% Labels
Random Initialization	41.8%	55.6%	63.4%	72.1%
Reconstruction Pretraining	50.7%	63.1%	69.8%	76.4%
Rotation Prediction	55.9%	68.7%	74.5%	79.6%

Contrastive Learning	63.8%	75.9%	80.7%	84.3%
Bootstrap Self-Supervision	65.4%	77.1%	81.9%	85.2%

6. Discussion

6.1 Pretext Tasks and Semantic Representation

The hypothetical comparison illustrates an important distinction between solving a self-supervised objective and learning features useful for later semantic recognition. Reconstruction encourages retention of detailed image information, but downstream classification may depend more strongly on invariance to nuisance factors. Earlier context prediction, jigsaw, colorization, and inpainting studies demonstrated that pretext tasks can produce transferable features when successful completion requires understanding meaningful image structure.

Transformation-prediction methods can also encourage semantic learning, but they may exploit unintended shortcuts when artificial transformations introduce recognizable artifacts. A rotation classifier, for example, may learn cues associated with image boundaries or texture orientation rather than higher-level objects. Effective self-supervision therefore depends on designing tasks in which the easiest successful solution requires the desired type of visual understanding. Downstream transfer evaluation provides an essential check against shortcut learning.

6.2 Contrastive and Bootstrap Representation Learning

The stronger hypothetical results assigned to contrastive learning reflect the ability of augmentation-based objectives to encode invariance across multiple views of the same image. SimCLR demonstrated the importance of augmentation composition and nonlinear projection, while MoCo showed how momentum-based dictionaries can support large collections of contrastive examples. Such methods directly optimize the geometry of representation space, making them especially suitable for downstream classification and retrieval.

Bootstrap approaches such as BYOL and SimSiam demonstrate that explicit negative examples are not always required. Their success also introduces a central methodological problem: networks trained only to make representations similar can potentially collapse to constant outputs. Different methods use momentum encoders, predictors, stop-gradient operations, clustering, or redundancy reduction to avoid this outcome. Understanding these mechanisms remains important when selecting a self-supervised strategy for new unlabeled image collections.

6.3 Label Efficiency and Transferability

The hypothetical limited-label results illustrate the practical motivation for self-supervised pretraining. If unlabeled images are abundant but annotations are expensive, useful representations can reduce the amount of labeled information required for a downstream classifier. This is particularly relevant for specialized visual domains in which annotation requires trained experts. However, benefits depend on similarity between pretraining data and the downstream visual problem.

Transferability should therefore be evaluated across more than one target task whenever possible. A representation that performs well for object classification may not provide equivalent advantages for segmentation, detection, medical imaging, or fine-grained recognition. MoCo and later self-supervised

methods have been evaluated across several downstream tasks, emphasizing that representation quality should be understood as a transfer property rather than performance on one benchmark alone.

7. Challenges

7.1 Data Augmentation and Semantic Preservation

Data augmentation plays a central role in modern self-supervised learning because it defines which transformations the representation should ignore. Cropping, color changes, grayscale conversion, and geometric transformations can encourage robustness to nuisance variability. However, excessive augmentation can remove information that is genuinely important to the downstream task. If color determines a class distinction, for example, aggressive color invariance may reduce useful semantic information.

The appropriate augmentation policy therefore depends partly on the structure of the target domain. Tian et al. demonstrated that view selection is a fundamental component of contrastive learning rather than a minor implementation detail. Empirical studies should report augmentation parameters carefully and conduct ablation analysis where possible. Improvements attributed to a new learning objective may otherwise result primarily from a stronger augmentation pipeline.

7.2 Representation Collapse and Optimization Stability

Representation collapse occurs when many different inputs are mapped to identical or nearly identical features, making the representation useless for discrimination. Contrastive methods avoid this outcome through negative examples, whereas non-contrastive methods employ alternative architectural or optimization mechanisms. BYOL uses an online and slowly updated target network, SimSiam emphasizes stop-gradient behavior, and Barlow Twins reduces redundancy through cross-correlation constraints.

These differences create practical implementation sensitivity. Batch size, normalization, momentum updates, projection dimensions, and optimization schedules can influence whether training remains stable. A comparison that provides substantially more tuning effort to one method may therefore be misleading. Self-supervised experiments should report optimization details and preferably repeat training across multiple random initializations before drawing conclusions regarding representation quality.

7.3 Computational Cost and Domain Generalization

Self-supervised pretraining can consume substantial computational resources because the network may process multiple transformed views of every image for many training epochs. Contrastive methods can additionally require large batches, memory queues, or multiple encoders. Although labeling requirements are reduced, computational cost does not disappear. Evaluation should therefore consider the performance improvement relative to pretraining time, memory usage, and hardware requirements.

Domain generalization creates an additional concern. Representations learned from natural everyday images may not transfer optimally to satellite, industrial, microscopic, or other specialized imagery. Domain-specific unlabeled pretraining may provide stronger features when sufficient in-domain images are available. Researchers should consequently distinguish general-purpose self-supervised pretraining from domain-adapted representation learning and evaluate transfer on data collected independently from the pretraining corpus.

8. Conclusion

Self-supervised learning provides a powerful methodological approach for extracting transferable visual representations from image collections that lack manual annotations. The present hypothetical study compared reconstruction, rotation prediction, contrastive learning, and bootstrap-based representation learning under a common framework. Representation quality was assessed using frozen-feature recognition and limited-label fine-tuning rather than the self-supervised training objective alone.

The illustrative results assigned the strongest transfer performance and label efficiency to contrastive and bootstrap-based strategies, while reconstruction and transformation prediction produced smaller hypothetical improvements. Every numerical result in this study is constructed and should not be interpreted as empirical evidence. Actual rankings can vary substantially according to architecture, augmentation, dataset structure, optimization resources, and the downstream task.

The study concludes that self-supervised learning is especially promising where unlabeled images are abundant and annotation is limited. Future empirical research should investigate augmentation design, collapse prevention, computational efficiency, cross-domain transfer, and robustness across multiple downstream tasks. Reliable comparison requires identical backbones, data splits, training resources, and transparent evaluation protocols so that improvements can be attributed to representation-learning strategy rather than experimental confounding.

References

1. Albelwi, S. (2022). Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4), Article 551. <https://doi.org/10.3390/e24040551>
2. Bao, H., Dong, L., & Wei, F. (2022). BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2106.08254>
3. Bardes, A., Ponce, J., & LeCun, Y. (2022). VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2105.04906>
4. Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 132–149). https://doi.org/10.1007/978-3-030-01264-9_9
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 9650–9660). <https://doi.org/10.1109/ICCV48922.2021.00951>
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 33. <https://arxiv.org/abs/2006.09882>
7. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607). PMLR. <https://proceedings.mlr.press/v119/chen20j.html>
8. Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15750–15758). <https://doi.org/10.1109/CVPR46437.2021.01549>

9. Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 1422–1430). <https://doi.org/10.1109/ICCV.2015.167>
10. Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. IEEE Signal Processing Magazine, 39(3), 42–62. <https://doi.org/10.1109/MSP.2021.3134634>
11. Fang, P.-F., Li, X., Yan, Y., Zhang, S., Kang, Q.-Y., Li, X.-F., & Lan, Z.-Z. (2022). Connecting the dots in self-supervised learning: A brief survey for beginners. Journal of Computer Science and Technology, 37, 507–526. <https://doi.org/10.1007/s11390-022-2158-x>
12. Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. In International Conference on Learning Representations (ICLR). <https://arxiv.org/abs/1803.07728>
13. Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., Kavukcuoglu, K., Munos, R., & Valko, M. (2020). Bootstrap your own latent—A new approach to self-supervised learning. In Advances in Neural Information Processing Systems, 33. <https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html>
14. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9729–9738). <https://doi.org/10.1109/CVPR42600.2020.00975>
15. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 16000–16009). <https://doi.org/10.1109/CVPR52688.2022.01553>
16. Jing, L., & Tian, Y. (2021). Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 4037–4058. <https://doi.org/10.1109/TPAMI.2020.2992393>
17. Misra, I., & van der Maaten, L. (2020). Self-supervised learning of pretext-invariant representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 6707–6717). https://openaccess.thecvf.com/content_CVPR_2020/html/Misra_Self-Supervised_Learning_of_P pretext-Invariant_Representations_CVPR_2020_paper.html
18. Noroozi, M., & Favaro, P. (2016). Unsupervised learning of visual representations by solving jigsaw puzzles. In Computer Vision—ECCV 2016 (pp. 69–84). Springer. https://doi.org/10.1007/978-3-319-46466-4_5
19. Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow Twins: Self-supervised learning via redundancy reduction. In Proceedings of the 38th International Conference on Machine Learning (pp. 12310–12320). PMLR. <https://proceedings.mlr.press/v139/zbontar21a.html>
20. Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In Computer Vision—ECCV 2016 (pp. 649–666). Springer. https://doi.org/10.1007/978-3-319-46487-9_40