Contrastive Self-Supervised Learning for Visual Recognition: A Survey

Ikboljon Sobirov
20020061@mbzuai.ac.ae

Muhammad Maaz
20020063@mbzuai.ac.ae

Roba Al Majzoub
20020071@mbzuai.ac.ae

1. Abstract

Self-supervised learning (SSL) has shown significant improvements in learning useful semantic representations from images. It has reduced the need for human-labeled data to acquire good performance in many applications. Specifically, SSL methods based on contrastive learning (CL) have been gaining popularity on account of their simplicity and exceptional results on many downstream tasks. In this work, we provide a thorough review of contrastive SSL techniques and focus further on neoteric methods. We discuss and compare the CL-based pipelines including MoCo, PIRL, SimCLR, BYOL, SwAV, SimSiam and Barlow Twins. SwAV utilizes online clustering and achieves 75.30% top-1 ImageNet linear classification accuracy, outperforming its counterparts. BYOL and Barlow Twins propose simple yet effective designs achieving 74.30% and 73.20% accuracies on the same downstream task. SimCLR-v2 achieves 77.50% top-1 accuracy using only 10% labels in the semi-supervised setting, surpassing supervised training by 1% margin. In addition, we discuss the importance of extending CL to make use of abundant uncurated image data. Specifically, we exemplify SEER that outperforms SwAV using SSL pre-training on one billion public Instagram images. Finally, we report the limitations of current approaches and argue that it is critical to design computationally efficient algorithms to increase the community participation.

Keywords: Contrastive Learning, Self-supervised Learning, Transfer Learning, Transform-Invariant Representations, Image Classification, Object Detection

2. Introduction

Deep learning models have achieved inspiring results in several fields [6, 13]. Although we humans can learn from a few samples, convolutional neural networks (CNNs) need to be trained on a huge amount of data to pull off a good accuracy. Therefore, the performance of these models substantially relies on the amount of labeled data available for training. However, preparing such large scale annotated data is time consuming and expensive.

Many self-supervised learning (SSL) routines have been proposed as a solution to mitigate such a challenging need for large scale labeled data [9, 18]. In these approaches, the pipeline mainly consists of two stages: pretext task and downstream task. In the pretext task, the model learns useful representations utilizing abundant unlabeled data. The downstream task involves transferring these representations to a real world application.

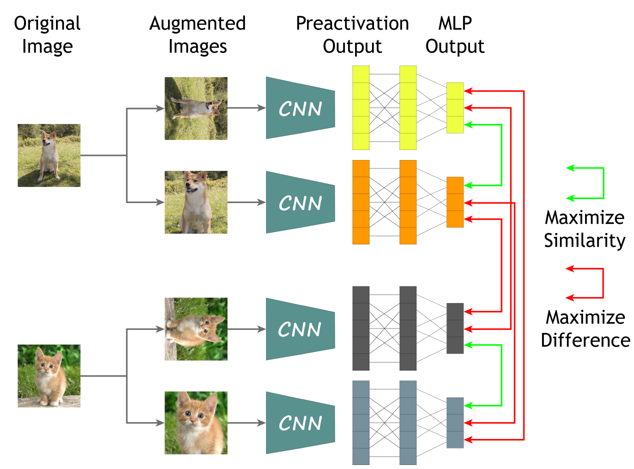

Figure 1: The figure shows a standard contrastive SSL pipeline. The CNN learns semantic features by maximizing the similarity between transformed views of the same image and maximizing the difference between transformed views of different images.

SSL pretext tasks can be generative (e.g. performing image inpainting), context based (e.g. solving jigsaw puzzle) or contrastive (e.g. making similar samples closer to and dissimilar samples farther from each other) [15, 14]. Contrastive learning (CL) tasks involve creating two different transformed views of the original image and maximizing the similarity between them. Concurrently, they involve maximizing the difference between transformed versions of different images, as shown in Fig. 1. In this survey, we scrutinize the recent advancements in CL techniques for visual recognition tasks.

Recently, CL has gained attention of the research community, and many papers have been published to manifest the potential of the technique [12, 17, 4, 20, 11]. MoCo [12] considers CL as an approach to train an encoder for a dictionary lookup task and outperforms supervised pre-training in several downstream tasks. PIRL [17] uses jigsaw to generate transformed views and utilizes a memory bank to surpass MoCo. Following these fast paced contributions, a few review articles have also appeared discussing many CL methods. However, these survey papers do not cover recent developments, such as Barlow Twins [22], SimTriplet [16] and SEER [10], to name a few. In this work, we focus on the novel CL methods and discuss possible future directions. Our contributions are as follows:

1. We provide a detailed analysis of new CL-based SSL methods necessary to understand the frontier concepts of the field.
2. We discuss scaling the CL-based SSL techniques on random and uncurated images. Note that this setting is closer to real world applications.
3. We analyze limitations of the techniques and propose potential future research directions.

3. Overview of Existing Methods

In a contrastive SSL approach, the model learns useful semantic features which are invariant to the distortions applied to the input images. Each image is transformed into two augmented views which are passed to an encoder network to generate vector representations. The contrastive loss function is defined to maximize the similarity between transformed versions of the same image and minimize the similarity between transformed versions of different images. However, the setting is pruned to learn a trivial constant representation, that is, converge to constant mapping for each input. Many different architectures have been proposed to avoid such a collapse and learn useful semantic features. In this section, we provide details about these approaches.

MoCo: Momentum Contrast (MoCo) [12] considers CL as a dictionary lookup task by introducing a moving average encoder to build dictionaries of the features of negative samples. The original image is transformed into two augmented views, called a query (x^{query}) and a key (x^{key}). The query is passed through a learnable encoder and the key is passed through a momentum encoder whose weights are updated as the moving average of the first encoder. It avoids the need for large batches by maintaining a dictionary of the features of negative samples. MoCo uses InfoNCE [21] as the contrastive loss (Fig. 2a).

As an update, the authors incorporate improvements from SimCLR [4] into MoCo and name it MOCO-v2. Specifically, they include an multi-layer perceptron (MLP) projection, extra augmentations and a cosine learning rate.

PIRL: Pretext-Invariant Representation Learning (PIRL) [17] demonstrates that it is critical to learn transformation-invariant representations since visual semantics are not affected by image transformations. It proposes asymmetric networks to avoid learning constant representations (Fig. 2b). The input image is passed through an encoder followed by a linear projection head that outputs first representation. Simultaneously, nine jigsaw [18] patches are created from the original image and are passed to the same encoder to generate feature vectors. These features are concatenated and passed through another linear projector to obtain a second representation. The noise contrastive estimator (NCE) computes the loss using these two representations as a positive pair and considers representations in the memory bank as negative samples. The memory bank enables PIRL to mitigate the need for large batches.

SimCLR: SimCLR [4] proposes a simple framework for contrastive learning of visual representations where a dedicated architecture or memory bank is not necessary (Fig. 2c). It performs a composition of augmentations in a batch of images, then passes them to an encoder followed by an MLP projection head. The paper uses an *NT-Xent* (normalized temperature-scaled cross entropy loss) loss defined as:

$$l_{i,j} = -log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_k)/\tau)} \tag{1}$$

where z_i, z_j are projection head outputs for the augmented views of same image, and τ is a temperature hyperparameter. Experiments show that the choice of augmentations broadens predictive tasks of the network and removes the constraints on the choice of its architecture. The paper also shows that a larger batch size and a higher number of training steps aid the learning process.

As an update, the authors present SimCLR-v2 [5] by introducing a deeper projection head along with the memory mechanism from MoCo [12] into SimCLR.

BYOL: Bootstrap Your Own Latent (BYOL) [11] uses two asymmetric networks: an online network and a target network, where a prediction head is only present in the online network (Fig. 2d). Two different augmented views of the same image are fed to online and and target networks. The online network predicts the representations of the target network and updates its parameters by minimizing the mean squared error (MSE) [19] between the two representations. The parameters of the target network are updated as the moving average of the parameters of the online network. The asymmetry between online and target networks, along with the use of moving average to update the weights of the target network, help BYOL to avoid learning collapsed representation. BYOL does not require negative samples; therefore, no large batches are needed during training.

SwAV: Swapping Assignments between Views (SwAV) [3] is an online clustering algorithm that takes advantage of CL without requiring to compute pairwise comparisons. Each input image is transformed into multiple augmented

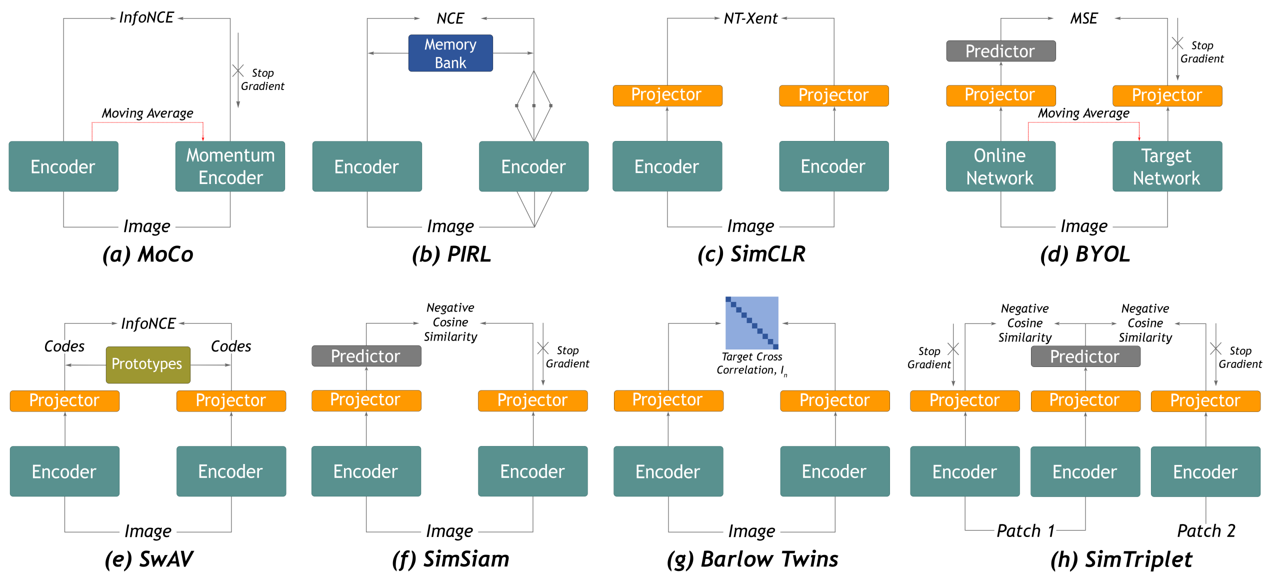

Figure 2: The figure shows different CL architectures reviewed in this paper. Projectors represent MLPs and encoders represent CNN backbones. MoCo and PIRL use linear projection heads which are not explicitly shown in the figure.

views that are mapped to features using symmetric network twins (Fig. 2e). The cluster assignments (i.e. codes) are calculated using these features. The algorithm enforces consistency between codes for different augmented views of the same image by predicting the codes of a view from the representation of another view. This is called a swap prediction loss and is given by Equation 2, where $l(z, q)$ measures the fit between features z and a code q.

$$L(z_t, z_s) = l(z_t, q_s) + l(z_s, q_t) \quad (2)$$

This method avoids trivial constant representations through online clustering and the swap prediction loss.

SimSiam: SimSiam [8] utilizes a simple Siamese network that maximizes the similarity between augmented views of the original image. It does not rely on negative samples as in SimCLR [4], momentum encoder as in MoCo [12] and BYOL [11] or online clustering as in SwAV [3] to avoid collapsing to a single constant representation. Rather, it uses a simple Siamese network with asymmetric network twins and stop gradient operations to learn good, semantically-meaningful representations (Fig. 2f). As it does not rely on negative samples, it does not require large batches during training. It minimizes the negative cosine similarity between the representations of augmented views of the same images during training.

Barlow Twins: Barlow Twins [22] proposes an objective function that naturally avoids trivial constant representation by measuring the cross-correlation matrix between the outputs from two similar networks and making it close to the identity matrix (Fig. 2g). The suggested objective function is given by Equation 3, where C is a square cross-correlation matrix calculated between the representations of the augmented views of the same image.

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1-\mathcal{C}_{ii})^2}_{\text{invariance term}} + \underbrace{\lambda \sum_i \sum_{j\neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}} \quad (3)$$

The invariance term in Equation 3 tries to equate the diagonal elements of C to one, making the representations invariant to the applied transformations. The redundancy reduction term pushes the off-diagonal elements to zero, decorrelating the different vector components of the representations and minimizing the redundancy between them. The concept is taken from the neuroscientist H. Barlow's redundancy-reduction principle [1] and is named as Barlow Twins.

SimTriplet: SimTriplet [16] applies contrastive SSL on medical images. It benefits from a multi-view nature of medical images and maximizes inter-patch and intra-patch similarity between them. Two neighboring patches are obtained from a medical Whole Slide Image (WSI [1]). Both patches undergo augmentations, and two views of the first patch along with one view of the second are fed simultaneously to the network. All three views are passed through identical encoders that output feature representations. The representation of one of the first patch's views is fed to the predictor. Similarity between the two representations and the output of the predictor is maximized by calculating the loss using a negative cosine distance (Fig. 2h). The method does not require negative samples, and thus call off the need

[1] https://www.mbfbioscience.com/whole-slide-imaging

Method	Momentum Encoder	Memory Bank	Projector Type	Predictor	Asymmetric Networks	Large Batches	Stop Gradient	Negative Pairs
MoCo	✓	✗	Linear	✗	✓	✗	✓	✓
PIRL	✗	✓	Linear	✗	✓	✗	✗	✓
SimCLR	✗	✗	MLP	✗	✗	✓	✗	✓
MoCo-v2	✓	✗	MLP	✗	✓	✗	✗	✓
SimCLR-v2	✓	✗	MLP	✗	✓	✓	✗	✓
BYOL	✓	✗	MLP	✓	✓	✗	✓	✗
SwAV	✗	✗	MLP	✗	✗	✗	✗	-
SimSiam	✗	✗	MLP	✓	✓	✗	✓	✗
Barlow Twins	✗	✗	MLP	✗	✗	✗	✗	✗
SimTriplet	✗	✗	MLP	✓	✓	✗	✓	✗

Table 1: The table shows a comparison of different contrastive SSL methods based on the architecture components used during the pretext task. Predictor heads used in these methods are MLPs.

for memory banks and large batches. The use of asymmetric networks along with stop-gradient allow the system to avoid trivial constant representations.

Table 1 compares the architectures of the discussed methods. Note that Barlow Twins [22] naturally avoids collapsed representations without requiring momentum encoder, memory bank, asymmetric networks or non-differentiable operations such as stop gradient.

4. Results

As aforesaid, the SSL pipeline consists of two, pretext and downstream, tasks. The knowledge learnt during the pretext task is employed in the downstream tasks that are application-oriented implementations, such as classification, object detection, segmentation, and so forth [15]. This section focuses on comparing how well the reviewed models perform for the downstream tasks.

ImageNet [6] and Places [23] are the top datasets for image classification. Pascal VOC [7] is the dataset commonly used for object detection. The analyzed papers use ResNet as a backbone for their encoders. Since all the papers follow a similar set of evaluation metrics, we will use these metrics to compare the results of the performances.

Linear evaluation on ImageNet, semi-supervised learning with ImageNet and transfer learning with VOC have been accepted by many as the common measurement protocols. In the linear evaluation method, the pre-trained networks are frozen and a linear classifier is learnt on top of them [15]. In the semi-supervised approach, a fraction of the whole data is sampled, where the whole network is fine-tuned without the need for regularization. The reviewed methods use 1 and 10 percent of ImageNet labeled data in a class-balanced way to show the significance of SSL [15]. Finally, the transfer learning uses the learnt representations to perform well on other downstream settings such as object detection [15].

Table 2 summarizes the results of the reviewed methods along with the results of a baseline supervised learning with ResNet50 for comparison. SwAV achieves a top performance of 75.3% for the ImageNet linear evaluation among the SSL models, which makes its performance significantly closer to that of supervised learning. It also shows the best performance for the object detection task with VOC07 and VOC07+12 datasets, scoring accuracies of 88.9% and 82.6% respectively. It is worth mentioning that, in both cases, the model is outperforming even the supervised baseline by 1.4% and 1.3% in a respective manner. The architecture of SwAV is fairly simple and similar to others, yet its performance is outstanding on account of online clustering.

BYOL, SwAV, SimSiam and Barlow Twins are comparable in that their architectures are somewhat similar, and therefore, achieve top results (74.3%, 75.3%, 71.3% and 73.2% respectively) in linear evaluation on ImageNet. The noticeable part is that these results are remarkably close to that of the baseline supervised network.

As for the semi-supervised learning with ImageNet, both in top-1 and top-5 accuracies, SimCLR-v2 tops the other methods. The results for 1% and 10% of the labeled data for the given methods are depicted in the Table 2. It is surprising that SimCLR-v2 slightly outperforms even the baseline supervised model in the top-1 accuracy with 77.5% compared to 76.5% using only 10% labels. In the top-5 accuracy, it tops again with the results of 91.5% and 93.4% using 1% and 10% of labels respectively. Knowledge distillation of SimCLR-v2 is assumed to be the main reason for such outstanding results. It may be achieving slightly less than the baseline supervised in the top-1 accuracy measurement with 1% labels, but it is fair to claim that a mere drop of 3.6% (77.5% to 73.9%) is acceptable when the annotation cost is reduced by 99%.

5. Discussion

Table 2 shows the results of linear evaluation, semi-supervised learning and transfer learning for different con-

Method	Top-1 Accuracy	Top-1 Accuracy		Top-5 Accuracy		Object Detection-mAP_{50}	
	Linear Evaluation	1% Labels	10% Labels	1% Labels	10% Labels	VOC07	VOC07+12
Supervised	76.50	25.40	56.40	48.40	80.40	87.50	81.30
MoCo	60.60	-	-	-	-	-	81.40
PIRL	63.60	30.70	60.40	57.20	83.80	81.10	80.70
SimCLR	69.30	48.30	65.60	75.50	87.80	85.50	-
MoCo v2	71.10	-	-	-	-	86.40	82.50
SimCLR v2	71.70	**73.90**	**77.50**	**91.50**	**93.40**	-	-
BYOL	74.30	53.20	68.80	78.40	89.00	85.40	-
SwAV	**75.30**	53.90	70.20	78.50	89.90	**88.90**	**82.60**
SimSiam	71.30	-	-	-	-	-	82.40
Barlow Twins	73.20	55.00	69.70	79.20	89.30	86.30	**82.60**

Table 2: The table shows the results of ImageNet linear evaluation, semi-supervised learning using 1% and 10% labels and transfer learning (TL) on Pascal VOC object detection (VOC07, VOC07+12) tasks. All trainings are initialized by the embeddings obtained from SSL pretraining on ImageNet. The embedding weights are frozen in ImageNet linear evaluation.

trastive SSL methods. Note that the ImageNet linear evaluation accuracy is not directly related to the transfer performance on the VOC detection task. For example, PIRL outperforms MoCo on ImageNet linear evaluation; however, MoCo performs better when transferred to the VOC07+12 detection task. Moreover, the performance gap for the object detection task between supervised and SSL methods is relatively large for small VOC07 dataset as compared to relatively large VOC07+12 dataset. This indicates that the features learned through SSL pipelines benefit more from large downstream task dataset. In addition, SwAV surpasses other CL-based SSL methods indicating the effectiveness of online clustering in learning good semantic features.

For semi-supervised benchmarks, SimCLR-v2 outperforms supervised training using only 10% of labels (77.50% vs 76.50%). It trains a big ResNet model (795M parameters) and distills knowledge to a smaller ResNet 50 model to achieve these results. This shows that wider and deeper self-supervised networks are powerful semi-supervised learners, and knowledge distillation is handy to prepare small deployable models achieving high performance.

Most previous works on SSL [18, 2, 5] focus on carefully chosen and processed data with predefined labels. These datasets represent only a fraction of the internet scale images. SEER [10] explores the potential of CL-based SSL methods in a setting closer to the real world by utilizing 1B random, uncurated public Instagram images for pretraining. It achieves similar or higher performance on different benchmarks compared to techniques as SimCLRv2 [5] and SwAV [3]. This indicates that the contrastive SSL methods can be extended to utilize abundantly available images and provide building blocks for online learning, that is, continue to improve the models as more data becomes available. The results are encouraging and further study is required to explore and design architectures that can benefit from copious image/video data uploaded on the web daily.

Current contrastive SSL methods require significant resources for training. For example, a single training for BYOL [11] and Barlow Twins [22] takes 113 and 124 hours respectively on 32 V100 GPUs. Therefore, only the substantial research groups such as FAIR [2] and Google Research [3] are actively contributing towards CL-based SSL methods. To recapitulate, it is critical to make contrastive SSL methods computationally efficient to make them accessible to a wide research community.

6. Conclusion

Self-supervised CL approaches have outperformed all previous techniques and reduced the need of large annotated data to achieve promising accuracies on many downstream tasks. The paper lays out details of contrastive SSL methods and emphasizes recently published approaches. There exists a trivial constant representation in a common contrastive SSL setting which is addressed by these new techniques. MoCo, PIRL, BYOL, SimSiam, and SimTriplet use asymmetric network twins to avoid this clash. SimCLR and SwAV use large negative samples and online clustering respectively as a solution. Barlow Twins naturally avoids learning collapsed representations using the redundancy reduction objective function. On ImageNet linear evaluation benchmarks, SwAV outperforms all other techniques indicating the usefulness of online clustering. In a semi-supervised setting, SimCLR-v2 surpasses supervised training by using only 10% labels, indicating that semi-supervised learning benefits more from big self-supervised models. We suggest exploring random uncurated images for CL-based SSL and developing computational and memory efficient pipelines as future research directions. This work will facilitate further research in the field by providing a comprehensive walk-through of the present methods.

[2] FaceBook AI Research: https://ai.facebook.com

[3] Google Research: https://ai.google/research

References

[1] H. Barlow. Redundancy reduction revisited. *Network (Bristol, England)*, 12:241–53, 09 2001. 3

[2] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, 2018. 5

[3] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. 2, 3, 5

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 1, 2, 3

[5] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020. 2, 5

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4

[7] M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, June 2010. 4

[8] L. Fan, H. Liu, and Y. Hou. An improved siamese network for face sketch recognition. In *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 1–7, 2019. 3

[9] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1

[10] P. Goyal, M. Caron, B. Lefaudeux, M. Xu, P. Wang, V. Pai, M. Singh, V. Liptchinsky, I. Misra, A. Joulin, and P. Bojanowski. Self-supervised pretraining of visual features in the wild, 2021. 2, 5

[11] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2, 3, 5

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 1, 2, 3

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. 1

[14] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1), 2021. 1

[15] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–1, may 5555. 1, 4

[16] Q. LIU, P. C. Louis, Y. Lu, A. Jha, M. Zhao, R. Deng, T. Yao, J. T. Roland, H. Yang, S. Zhao, L. Wheless, and Y. Huo. Simtriplet: Simple triplet representation learning with a single gpu. *ArXiv*, abs/2103.05585, 2021. 2, 3

[17] I. Misra and L. van der Maaten. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716, 2020. 1, 2

[18] M. Noroozi and P. Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1, 2, 5

[19] C. Sammut and G. I. Webb, editors. *Mean Squared Error*, pages 653–653. Springer US, Boston, MA, 2010. 2

[20] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020. 1

[21] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 2

[22] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021. 2, 3, 4, 5

[23] B. Zhou, Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems*, 1, 05 2015. 4