Large-Scale Hyperspectral Image Clustering Using Contrastive Learning

Yaoming Cai[1], Yan Liu[1], Zijia Zhang[1], Zhihua Cai[1,4] and Xiaobo Liu[2,3]

[1]School of Computer Science, China University of Geosciences, 430074 Wuhan, China
[2]School of Automation, China University of Geosciences, 430074 Wuhan, China
[3]Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, China
[4]Corresponding author

Abstract

Unsupervised hyperspectral image (HSI) classification is an important but challenging task in the hyperspectral processing community. Despite great success, previous HSI clustering approaches belong to offline clustering which is often performed in a transductive scheme, thus failing to generalize to large-scale and unseen scenes. In this paper, we propose an online and deep clustering model for large-scale HSI clustering, termed spectral-spatial contrastive clustering (SSCC). Specifically, SSCC performs contrastive learning based on a series of semantic-preserving spectral-spatial augmentation to simultaneously maximize the intra-class agreement and inter-class variation, which are implemented by an instance-level contrastive loss and a cluster-level contrastive loss, respectively. The SSCC model is trained in an end-to-end fashion with minibatch, allowing it to efficiently handle large-scale HSIs. We assess the performance of SSCC on real HSI and show that SSCC significantly advances the state-of-the-art results with 8.41% improvement on accuracy.

Keywords

Hyperspectral image, clustering, self-supervised learning, contrastive learning

1. Introduction

Hyperspectral image (HSI) consists of hundreds of narrow bands with rich spectral and spatial information, revealing the spectral property of the area or object of interest at a nanometer resolution [1]. HSI intelligent interpretation is one of the hot spots in the current remote sensing community. With the development of deep learning techniques, great progress has been made by training expressive networks with massive labelled data [2]. However, current human-annotated datasets suffer from a large amount of manpower, leading to the limitation of availability and applicability [3].

Without label information, unsupervised HSI classification becomes a challenging task, thus leading to uncompetitive accuracy. Many efforts have been devoted to bridging the gap between supervised models and unsupervised models [4]. More recently, subspace clustering (SC) [5, 6] and non-negative matrix factorization (NNMF) [7, 8] were frequently adopted for HSI clustering. Despite their promising performance, these approaches collectively suffer from two drawbacks. First, they are based on shallow feature representation and failing to capture high-level spectral-spatial information, which results in poorer robustness and generalization ability. Second, they focus on offline HSI clustering tasks, i.e., the clustering is dependent upon the whole dataset, which limits their application on large-scale online learning scenarios.

To address the first drawback, some attempts, e.g., [3, 9, 10], have been made to use deep clustering networks to learn cluster-friendly deep representations. Such approaches can usually improve the clustering performance upon the shallow ones by significant margins. However, the second drawback remains an open problem. Zhai et al. [11] shown that sparsity representation is useful to alleviate the issue. Nonetheless, the procedure of constructing a suitable dictionary if often heuristic and suboptimal, particularly cannot be implemented by end-to-end. As a result, most current works compromise on this problem by verifying on smaller scenes, lacking dependable performance evidence from large-scale HSI data.

Fortunately, self-supervised learning (SSL) has emerged as a powerful paradigm to circumvent human annotation [12]. The core idea is to learn to solve a label-free pretext task, such as colorization [13] and inpainting [14], enabling the model to capture semantic information. A downstream task will benefit from the pre-trained model by fine-tuning and transfer learning. According to their objectives, pretext tasks can be broadly classed into three categories [15]: generative, contrastive, and adversarial. The tremendous success of recent contrastive learning models including SimCLR [16], BYOL [17], MoCo [18], and BalowTwins [19], has proven that contrastive learning tends to be a more promising branch. The pretext in contrastive learning is to maximize the similarity between two positive views of every sample that are auto-

CDCEO 2021: 1st Workshop on Complex Data Challenges in Earth Observation, November 1, 2021, Virtual Event, QLD, Australia.
caiyaom@cug.edu.cn (Y. Cai); yanliu@cug.edu.cn (Y. Liu); zhangzijia@cug.edu.cn (Z. Zhang); zhcai@cug.edu.cn (Z. Cai); xbliu@cug.edu.cn (X. Liu)

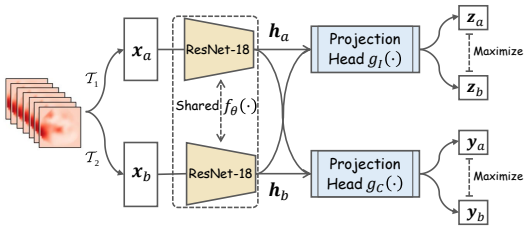


Figure 1: The overall framework of our proposed SSCC. Two augmentations are sampled from an augmentation pool $\mathcal{T}$ and applied to input patches. A shared backbone encoder $f(\cdot)$ and two projection heads, i.e., instance-level projection head $g_I(\cdot)$ and cluster-level projection head $g_C(\cdot)$, are trained to simultaneously maximize the agreement between instance representations $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$, and cluster representations $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$ via a contrastive loss.

matically generated by data augmentation.

In this paper, we propose a spectral-spatial contrastive clustering (SSCC) approach for large-scale HSI. The approach takes ResNet-18 as the backbone and consists of an instance-level contrastive head and a cluster-level contrastive head. Considering the inherent spectral-spatial properties of HSI, we introduce several semantic-preserved augmentation strategies, including ResizedCrop, Horizontal/Vertical Flip, and GroupBandShuffle. The proposed approach has some unique advantages: 1) SSCC performs clustering and deep feature learning simultaneously; 2) SSCC adopts minibatch training in an end-to-end fashion, thus it is inherently suitable for large-scale HSI scenes; 3) SSCC is an online clustering model and can easily generalize to unseen data.

2. Methodology

Motivated by recent contrastive clustering developments in visual representation learning [20], which performs clustering jointly with contrastive learning, we introduce a novel SSCC approach for large-scale HSI clustering.

2.1. Overall

The core of SSCC is to maximize the similarity between representations of positive pairs from both instance space and cluster space, as shown in Fig. 1. The SSCC conducts a spectral-spatial augmentations, then the augmented pairs are forwarded into a weight-sharing backbone encoder, $f(\cdot)$, resulting in deep representations $\boldsymbol{h}_a$ and $\boldsymbol{h}_b$. Behind the encoder, projection heads consisting of instance projection head and cluster projection head are carried out to maximize the similarity between prediction pairs. More specifically, we adopt ResNet-18 and MLPs as the backbone and the projection heads respectively, in which the instance/cluster projection head transforms data into 128 and C dimension, where C denotes the number of targets. We use the cluster projection head to perform clustering at the inference stage.

2.2. Spectral-Spatial Augmentation

Formally, let $\boldsymbol{x}$ be an HSI sample in $\mathbb{R}^{n_1 \times n_2 \times m}$, where $n_1 \times n_2$ is the spatial size and m denotes the number of spectral band. We construct positive pair by forwarding $\boldsymbol{x}$ to two augmentations $\mathcal{T}_a$ and $\mathcal{T}_b$ sampling from an augmentation pool $\mathcal{T}$. Formally, $\boldsymbol{x}_a = \mathcal{T}_a(\boldsymbol{x})$ and $\boldsymbol{x}_b = \mathcal{T}_b(\boldsymbol{x})$, where $\mathcal{T}_a, \mathcal{T}_b \in \mathcal{T}$.

Based on the characteristics of HSI, the augmentation pool consists of spectral augmentations and spatial augmentations. The spectral augmentations include group band permutation and band random drop, and the spatial augmentations include random crop with resize and random horizontal/vertical flip. Precisely, group band permutation divides m bands into k adjacent groups and randomly permutes spectral bands within each group. Band random drop will mask a spectral band with a probability of p. The spatial augmentations are the same as the pipelines defined in torchvision[1].

2.3. Projection Heads

SSCC contains two projection heads. We use $g_I(\cdot)$ and $g_C(\cdot)$ to denote the instance-level projection head and cluster-level projection head. Each head takes $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$ as inputs and produces a pair of predictions, i.e., denoting as $\boldsymbol{z}_a$ and $\boldsymbol{z}_b$ for $g_I(\cdot)$ and $\boldsymbol{y}_a$ and $\boldsymbol{y}_b$ for $g_C(\cdot)$. The goal of $g_I(\cdot)$ is to encourage the intra-class agreement, instead and $g_C(\cdot)$ aims to encourage the inter-class variation. Specifically, we achieve these by defining the following contrastive losses. Let $\{\boldsymbol{x}_a^{(1)}, \ldots, \boldsymbol{x}_a^{(N)}, \boldsymbol{x}_b^{(N+1)}, \ldots, \boldsymbol{x}_b^{(2N)}\}$ be $2N$ augmented samples with batch size of N. The instance-level contrastive loss over sample $\boldsymbol{x}_i^a$ is given by

$$\mathcal{L}_a^{(i)} = -\log\left(\frac{\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{z}_a^{(i)}, \boldsymbol{z}_b^{(i)}\right)}{\mathcal{T}_I}\right)}{\sum_{j=1}^{N}\left[\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{z}_a^{(i)}, \boldsymbol{z}_a^{(j)}\right)}{\mathcal{T}_I}\right)+\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{z}_a^{(i)}, \boldsymbol{z}_b^{(j)}\right)}{\mathcal{T}_I}\right)\right]}\right) \tag{1}$$

Here, $\mathcal{T}_I$ denotes a temperature parameter and sim is a similarity function. We adopt cosine similarity in the paper, i.e.,

$$\mathrm{sim}\left(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(j)}\right) = \frac{\left(\boldsymbol{z}^{(i)}\right)\left(\boldsymbol{z}^{(j)}\right)^T}{\|\boldsymbol{z}^{(i)}\|\,\|\boldsymbol{z}^{(j)}\|} \tag{2}$$

[1] https://pytorch.org/

Similarly, we calculate the loss of $\boldsymbol{x}_b^{(i)}$ by $\mathcal{L}_b^{(i)}$. The batched instance-level contrastive loss is defined as $\mathcal{L}_I = \frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}_a^{(i)} + \mathcal{L}_b^{(i)}$.

Instead, the cluster-level contrastive loss is defined on an inter-class space, i.e.,

$$\hat{\mathcal{L}}_a^{(i)} = -\log\left(\frac{\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{y}_a^{(\cdot i)},\boldsymbol{y}_b^{(\cdot i)}\right)}{\mathcal{T}_C}\right)}{\sum_{j=1}^{C}\left[\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{y}_a^{(\cdot i)},\boldsymbol{y}_a^{(\cdot j)}\right)}{\mathcal{T}_C}\right)+\exp\left(\frac{\mathrm{sim}\left(\boldsymbol{y}_a^{(\cdot i)},\boldsymbol{y}_b^{(\cdot j)}\right)}{\mathcal{T}_C}\right)\right]}\right) \tag{3}$$

where $\mathcal{T}_C$ is another temperature parameter. Furthermore, the cluster-level contrastive loss can be defined as

$$\mathcal{L}_C = \frac{1}{2C}\sum_{i=1}^{C}\left(\hat{\mathcal{L}}_a^{(i)} + \hat{\mathcal{L}}_b^{(i)}\right) - H\left(Y\right), \tag{4}$$

where $H\left(Y\right)$ denotes the entropy of cluster assignment probabilities across the whole augmented minibatch, which is used to avoid the trivial solution, and can be computed by

$$H\left(Y\right) = \sum_{i=1}^{C} P\left(\boldsymbol{y}_a^{(\cdot i)}\right)\log\left(P\left(\boldsymbol{y}_a^{(\cdot i)}\right)\right) + P\left(\boldsymbol{y}_b^{(\cdot i)}\right)\log\left(P\left(\boldsymbol{y}_b^{(\cdot i)}\right)\right). \tag{5}$$

Finally, the complete training loss function of SSCC is indicated as

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_C. \tag{6}$$

2.4. Training and Predicting

The proposed SSCC can be trained in an end-to-end fashion. Specifically, we adopt the widely-used Adam approach as the optimizer with a learning rate of 0.00002, batch size of 128, and L_2 regularizer of 0.00005. During the inference stage, we feedforward any given sample and lock the spectral-spatial augmentation process, the output of the cluster-level projection head is regarded as the prediction of the sample.

3. Experiments

3.1. Datasets and Setup

In this paper, we conduct experiments on the widely-used Indian Pines dataset. We follow the training settings suggested in [20] and the baselines reported in [11]. Several clustering metrics are used to quantize the clustering performance, including producer's accuracy, overall accuracy (OA), Kappa coefficient (Kappa), and purity. It should be noted that more datasets and more extensive experiments will be provided in our future work.

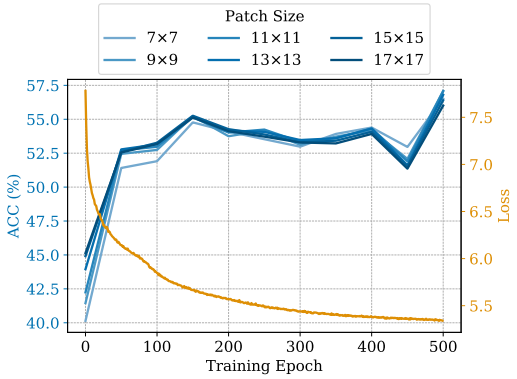


Figure 2: Loss/ACCs curves along with model training, where ACC is obtained under varying patch size.

3.2. Comparisons with State of the Arts

Table 1 reports the comparative results of different HSI clustering approaches. We can see that SSCC achieves state-of-the-art clustering results in terms of OA, Kappa, and purity. In particular, SSCC (OA=57.07) improves upon JSSC (OA=48.66) by a margin of 8.41 points. Furthermore, SSCC accurately distinguishes class Nos. 4, 8, 13, and 15, which is remarkably better than other baselines. This demonstrates that SSL-based HSI clustering not only has obvious theoretical edges but also has significant practical effectiveness.

3.3. Ablation Studies

In Fig. 2, we show the effect of patch size and training epoch. From the curves, we can conclude that: 1) The clustering ACC of SSCC is dramatically increased along with the training; 2) SSCC achieves considerable clustering ACC at a completely random initial status (epoch=0), signifying the feature representation power of SSL; 3) A larger patch size is often more beneficial to the SSCC model, especially when using a small train epoch.

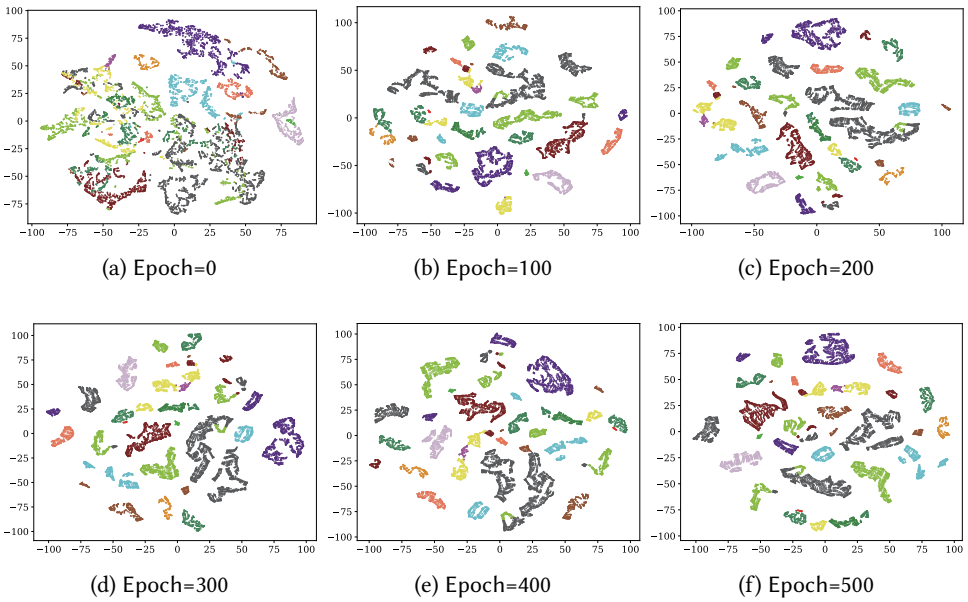
We further present the evolution of feature representations across the training process of SSCC, as shown in Fig. 3. It can be seen that the features tend to become more compact within a certain class and more separable from each other class. This proves that our SSCC can capture the intrinsic spectral-spatial information of HSI and obtain superior clustering performance and generalization ability.

4. Conclusion

This paper presented a novel SSCC model for large-scale HSI online clustering task. The SSCC model follows a

Table 1
Comparative results on the Indian Pines dataset.

Class No.	FCM	FCM-S1	SSC-S	L2-SSC	LRSC	SGCNR	FSCAG	SCC	JSCC	SSCC
1	39.13	15.22	0	0	52.17	26.52	9.57	8.26	13.48	0
2	25.63	26.58	26.75	42.58	27.31	36.13	23.85	27.87	33.40	29.48
3	24.58	26.10	36.87	18.80	1.81	22.80	34.36	41.47	30.17	57.71
4	6.33	13.92	10.13	0	13.50	13.00	24.47	9.62	16.12	100
5	44.31	46.29	62.73	62.32	57.97	44.68	40.08	46.00	57.93	65.84
6	26.99	29.78	72.05	80.41	32.47	44.08	39.45	52.41	55.97	95.34
7	0	0	0	89.29	0	14.29	0	12.86	0.71	0
8	86.61	98.83	100	54.39	28.03	66.78	92.59	78.62	72.51	100
9	20.00	29.00	0	25.00	25.00	12.00	9.00	5.00	15.00	0
10	23.15	25.60	35.08	46.81	17.18	34.75	29.12	27.65	48.66	48.05
11	28.19	26.66	37.68	37.19	69.86	34.60	30.22	39.41	58.71	34.22
12	23.61	24.72	30.69	31.53	19.90	15.82	22.56	12.65	19.26	75.04
13	99.02	98.73	99.02	95.61	23.90	74.54	87.71	87.12	61.27	100
14	34.16	32.98	49.33	45.69	41.03	44.36	37.45	38.62	70.31	69.33
15	17.62	18.81	15.54	15.28	19.17	15.60	17.67	19.07	25.49	100
16	59.14	58.06	97.85	94.62	0	59.14	65.81	69.03	37.63	0
OA(%)	31.35	32.70	43.37	43.11	36.68	36.31	34.70	37.76	48.66	**57.07**
Kappa	0.2561	0.2695	0.3757	0.3667	0.2713	0.2946	0.2887	0.3091	0.4254	**0.5390**
Purity	0.5015	0.5082	0.5588	0.5670	0.4571	0.5105	0.5137	0.5222	0.5689	**0.7475**



(a) Epoch=0 (b) Epoch=100 (c) Epoch=200

(d) Epoch=300 (e) Epoch=400 (f) Epoch=500

Figure 3: The evolution of feature representations across the training process, where the features for t-SNE are computed from the backbone.

contrastive learning pipeline and consists of two projection heads associated with the instance-level and the cluster-level contrasting. Furthermore, we introduced a semantics-preserving augmentation pool based on the characteristic of HSI. SSCC is featured by offline clustering, minibatch and end-to-end training, making it easy to deal with large-scale HSI. Experimental results on real HSI show that SSCC can achieve state-of-the-art clustering performance with significant margins over previous works. The success of SSCC offers a powerful alternative for unsupervised HSI classification. It should be noted that this is a preliminary work and further analysis on the proposed method will be conducted in our future works.

5. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61773355 and 61973285, the Fundamental Research Funds for National Universities, China University of Geosciences (Wuhan) under Grant CUGGC03 and 1910491T06, and the National Scholarship for Building High Level Universities, China Scholarship Council (No. 202006410044).

References

[1] Y. Cai, X. Liu, Z. Cai, Bs-nets: An end-to-end framework for band selection of hyperspectral image, IEEE Transactions on Geoscience and Remote Sensing 58 (2020) 1969–1984.

[2] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Hofle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. M. Atkinson, J. A. Benediktsson, Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art, IEEE Geoscience and Remote Sensing Magazine 7 (2019) 6–39.

[3] Y. Cai, Z. Zhang, Z. Cai, X. Liu, X. Jiang, Q. Yan, Graph convolutional subspace clustering: A robust subspace clustering framework for hyperspectral image, IEEE Transactions on Geoscience and Remote Sensing 59 (2021) 4191–4202.

[4] H. Zhai, H. Zhang, L. Pingxiang, L. Zhang, Hyperspectral image clustering: Current achievements and future lines, IEEE Geoscience and Remote Sensing Magazine (2021).

[5] H. Zhang, H. Zhai, L. Zhang, P. Li, Spectral-spatial sparse subspace clustering for hyperspectral remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 54 (2016) 3672–3684.

[6] H. Zhai, H. Zhang, L. Zhang, P. Li, Nonlocal means regularized sketched reweighted sparse and low-rank subspace clustering for large hyperspectral images, IEEE Transactions on Geoscience and Remote Sensing 59 (2021) 4164–4178.

[7] L. Zhang, L. Zhang, B. Du, J. You, D. Tao, Hyperspectral image unsupervised classification by robust manifold matrix factorization, Information Sciences 485 (2019) 154 – 169.

[8] Y. Qin, B. Li, W. Ni, S. Quan, P. Wang, H. Bian, Affinity matrix learning via nonnegative matrix factorization for hyperspectral imagery clustering, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14 (2021) 402–415.

[9] Y. Cai, M. Zeng, Z. Cai, X. Liu, Z. Zhang, Graph regularized residual subspace clustering network for hyperspectral image clustering, Information Sciences 578 (2021) 85–101.

[10] J. Lei, X. Li, B. Peng, L. Fang, N. Ling, Q. Huang, Deep spatial-spectral subspace clustering for hyperspectral image, IEEE Transactions on Circuits and Systems for Video Technology 31 (2021) 2686–2697.

[11] H. Zhai, H. Zhang, L. Zhang, P. Li, Sparsity-based clustering for large hyperspectral remote sensing images, IEEE Transactions on Geoscience and Remote Sensing (2020) 1–15. doi:10.1109/TGRS.2020.3032427.

[12] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1. doi:10.1109/TPAMI.2020.2992393.

[13] R. Zhang, P. Isola, A. A. Efros, Colorful image colorization, in: European conference on computer vision, Springer, 2016, pp. 649–666.

[14] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A. A. Efros, Context encoders: Feature learning by inpainting, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2536–2544.

[15] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, J. Tang, Self-supervised learning: Generative or contrastive, IEEE Transactions on Knowledge and Data Engineering (2021) 1–1. doi:10.1109/TKDE.2021.3090866.

[16] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 1597–1607.

[17] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, M. Valko, Bootstrap Your Own Latent: A new approach to self-supervised learning, in: Neural Information Processing Systems, Montréal, Canada, 2020.

[18] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[19] J. Zbontar, L. Jing, I. Misra, Y. LeCun, S. Deny, Barlow twins: Self-supervised learning via redundancy reduction, in: Proceedings of the 38th International Conference on Machine Learning, 2021.

[20] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, X. Peng, Contrastive clustering, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, AAAI Press, 2021, pp. 8547–8555.