SkiLL: Skipping Color and Label Landscape: Self Supervised Design Representations for Products in E-commerce

Vinay K Verma, Dween Rabius Sanny, Shreyas Sunil Kulkarni, Prateek Sircar
Abhishek Singh and Deepak Gupta
International Machine Learning, Amazon
vinayugc@gmail.com, drsanny@amazon.com, sskshreyas@gmail.com, sircarp@amazon.com
p15abhisheks@iima.ac.in and deepakgupta.cbs@gmail.com

Abstract

Understanding the design of a product without human supervision is a crucial task for e-commerce services. Such a capability can help in multiple downstream e-commerce tasks like product recommendations, design trend analysis, image-based search, and visual information retrieval, etc. For this task, getting fine-grain label data is costly and not scalable for the e-commerce product. In this paper, we leverage knowledge distillation based self-supervised learning (SSL) approach to learn design representations. These representations do not require human annotation for training and focus on only design related attributes of a product and ignore attributes like color, orientation, etc. We propose a global and task specific local augmentation space which captures the desired image information and provides robust visual embedding. We evaluated our model for the three highly diverse datasets, and also propose and measure a quantitative metric to evaluate the model's color invariant feature learning ability. In all scenarios, our proposed approach outperforms the recent SSL model by upto 8.6% in terms of accuracy.

1. Introduction

The automatic identification of visual properties, such as texture, pattern, style, and shape, is highly critical for e-commerce stores. This capability has many practical applications, including simulating an offline shopping experience for customers, enabling sellers and brands to understand and respond to current trends, and aiding in the recommendation of relevant products to customers. The visual feature-based identification should not be limited to clothing or home decor items, but should widely apply to a range of products, including fashion, electronics, furniture, beauty, and bags. One primary challenge is capturing the pattern, design, and shape while remaining agnostic to other properties, such as color and image orientation (e.g. image 2).

An usual e-commerce store may have millions of images

Figure 1. Same pattern label (striped) but visually different pattern

Figure 2. Similar design but different orientation and color

Figure 3. An example of 2 different products with similar pattern

for each product type, but labelling them is extremely challenging, costly and time consuming. Sometimes, sellers or brands provide the product attribute (e.g. 'floral' or 'solid' patterns) but these pattern labels are often too generic and coarse-grained, resulting in products with the same label (e.g., 'striped') that may not necessarily "look similar" in terms of pattern (as depicted in Figure 1). Self-supervised learning (SSL) has recently gained significant attention due to its ability to learn visual embeddings without labeled data [4, 5, 8, 13, 14, 21]. The SSL approaches, such as SimCLR [5], MoCo [17], and MoCoV2 [6], use image augmentation techniques (such as Gaussian Blur, color jittering, and random cropping) to generate positive pairs of the same image to learn visual similarity embeddings. However, these approaches also require negative samples for contrastive loss which is error prone. Recent works, BYOL [14], DINO [4], and SimSiam [8] have introduced self-knowledge distillation-based architectures that eliminate the use of negative pairs.

In the proposed model, we leverage the SSL approach to learn the desired feature. However, we observed that SSL has several shortcomings when used directly. Specifically, it fails to learn fine-grained design representations and is not agnostic to visual properties like color. To address these

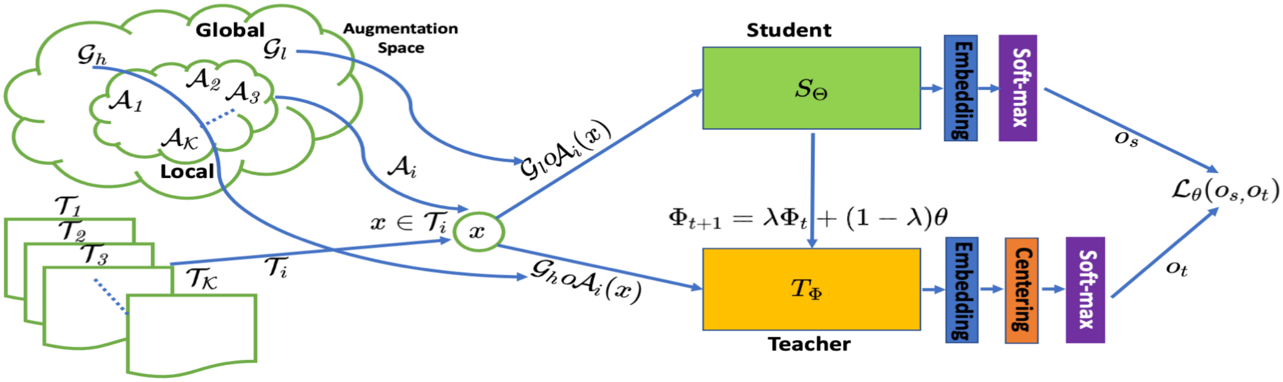

Figure 4. The block diagram of the proposed model, augmentation space contains the local and global augmentation. The local augmentation are selected based on the product type $\mathcal{T}_i$ and $\mathcal{G}_l$ and $\mathcal{G}_h$ augmentation are applied to achieve two pair.

shortcomings, we used additional augmentations to learn the task specific and color invariant feature. The e-commerce products are diverse and same augmentation does not work well across multiple product therefore we propose task specific augmentation. The proposed model is robust and shows the promising result for a diverse set of task as compared to vanilla SSL model. Also we propose a new metric to measure the model's color invariant learning ability. Our model has been extensively tested on diverse datasets, e.g. women's dresses, furniture, and shirts, demonstrating its effectiveness. On tested classification tasks, our models show increase by upto 8.6% in accuracy.

2. Related Work

The goal of this research is to develop a model that learns an embedding, such that images with similar visual properties are closer in the embedding space, also the embedding should be agnostic to the color [1, 11]. To address this problem, several proposals [3, 15, 19, 20, 22] have been made. Recently, self-supervised approach allows learning without the need for supervisory signals, has gained significant attention due to its ability to eliminate the costly annotations. The initial SSL work RotNet [13], CFN [21] shows good performance, however the recent work in the SSL focus on the contrastive [16] or knowledge distillation based learning and shows the promising performance. The approach [5, 7] leverages the contrastive learning [10, 16] between the positive and negative pair. These pairs are made using the augmentation and random sampling respectively. Here sampling the negative data is error prone therefore model achieves degraded results.

Another set of recent research BOYL [14], SimSiam [8], DINO [4] discard the requirement of negative samples and trains the model only using the original sample and its augmentation as a positive samples. BOYL and Simsiam model use CNN architecture as backbone and tries to learn the similar embedding of the augmented version of the same samples. Other work DINO [4] incorporates the transformer architecture as backbone and minimize the cross entropy loss between the two embedding of the teacher and student model. In these approach the model collapse is the key bottleneck, to overcome the above challenge knowledge distillation [2, 18] methods are used. During the distillation only one network [12] is trained and other learns using the exponential moving average.

3. Proposed Method

The e-commerce products encompass a wide range of diverse categories, including apparel (e.g. dresses, shirts, pants, jackets), curtains, furniture, toys, wall art, and shoes. Gathering labeled data for these categories can be a challenging and expensive process. The proposed approach is based on the self-supervised learning approaches [4, 5, 8, 14]. Similar to DINO [4] our approach leverages the vision transformer [9] as the base architecture for image encoding. The fixed augmentation technique does not work well for the diverse category since some key attribute in a category may be not required for others, also in the recommendation of design or pattern color are not important therefore model should learn the color invariant feature. To achieve this, we propose a data augmentation technique that captures the desired information and produces the robust output.

3.1. Notations

We consider a set of tasks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$, where each task corresponds to a specific product type. Let (S_θ) be the student model and (T_ϕ) be the teacher model, where θ and ϕ are the parameters of the student and teacher, respectively.

3.2. Augmentation Space

Data augmentation is crucial in SSL models, and can greatly impact the performance of the model. In our approach, we define two types of augmentations: global augmentations, which capture generic information that is common across tasks, and local augmentations, which capture task-specific information. Let $\mathcal{G} = \mathcal{G}_h, \mathcal{G}_l$ represent global augmentations and let $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \ldots, \mathcal{A}_K$ represent local augmentations. Here, $\mathcal{G}_{\{h,l\}} = g_1, g_2, \ldots g_g$ and $\mathcal{A}_i = a_1, a_2, \ldots a_l$, where $a_i \in \mathcal{A}$ and $\mathcal{A}$ represents the local augmentation space. Each local $\mathcal{A}_i$ contains a set of augmentations from the local augmentation space. During the training of the i^{th} task $\mathcal{T}_i$, we select the global augmentation

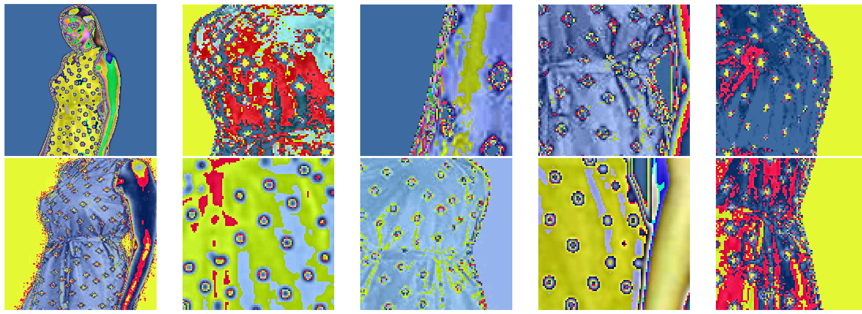

Figure 5. Global augmentation space, $\mathcal{G}_h$ (first column), $\mathcal{G}_l$ (remaining). Here $\mathcal{G}_h$ are of high resolution (224×224) and $\mathcal{G}_l$ are of low resolution (96×96).

Figure 6. Local augmentation space, $\mathcal{A}_i$ for the women-dress.

$\mathcal{G}_h, \mathcal{G}_l$ and the corresponding local augmentation $\mathcal{A}_i$. More details around the augmentations is in the supplementary sheet. Also for a product, in e-commerce, there are typically multiple images (5-8) of a single product. Let $\boldsymbol{x}'$ and $\boldsymbol{x}''$ be two random images from the same product. The positive pair for the $\boldsymbol{x}'\boldsymbol{x}'' \in \mathcal{T}_i$ is defined as follows:

$$\boldsymbol{x}_1, \boldsymbol{x}_2 = \mathcal{G}_h o \mathcal{A}_i(\boldsymbol{x}'), \mathcal{G}_h o \mathcal{A}_i(\boldsymbol{x}'') \tag{1}$$

$$\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_c = \mathcal{G}_l o \mathcal{A}_i(\boldsymbol{x}') \tag{2}$$

In our experiments, $\mathcal{G}_h$ and $\mathcal{G}_l$ are high and low-resolution images. $\mathcal{G}_h$ generates two positive pairs and $\mathcal{G}_l$ generates c positive pairs. o is a composite operation. $\mathcal{G}_h o \mathcal{A}_i(\boldsymbol{x})$ and $\mathcal{G}_l o \mathcal{A}_i(\boldsymbol{x})$ give the output after the composite operation. From the set of local augmentations, we can choose any task-specific augmentation and find images with positive pairs $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. See Figure 4, 5, 6 for a visual illustration.

3.3. Teacher-Student framework for SSL

We have a set of tasks $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_K$ where each task is a different product type. During the training of the i^{th} task $\mathcal{T}_i$, we select $\mathcal{G}_h$ and $\mathcal{G}_l$ as the global and $\mathcal{A}_i$ as the local augmentation. Say, after applying the augmentation ($\mathcal{G}_h o \mathcal{A}_i(\boldsymbol{x})$) we get $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which are the high resolution augmented positive pair and ($\mathcal{G}_l o \mathcal{A}_i(\boldsymbol{x})$) gives $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_c$ which are low resolution positive pair. The augmentation obtained by $\mathcal{G}_h$ and $\mathcal{G}_l$ passed to the network T_ϕ and S_θ respectively, which is given as:

$$\boldsymbol{o}_{s_i} = SM(S_\theta(\boldsymbol{x}_i)); \quad \& \quad \boldsymbol{o}_{t_i} = SM(C_e(T_\phi(\boldsymbol{y}_i))) \tag{3}$$

Where C_e is the centering function and SM is the Softmax operation. The Softmax operation for the input $S_\theta(\boldsymbol{x}_i)$ is:

$$\boldsymbol{o}_{s_i} = \frac{\exp(S_\theta(\boldsymbol{x}_i)/\tau)}{\sum_{i=1}^{N} \exp(S_\theta(\boldsymbol{x}_i)/\tau)} \tag{4}$$

Here τ is the temperature parameter which is used to flatten the output space. The loss between the output $\boldsymbol{o}_{s_i}$ and $\boldsymbol{o}_{t_i}$ is given as: $\mathcal{L}(\boldsymbol{o}_{s_i}, \boldsymbol{o}_{t_i}) = -\boldsymbol{o}_{t_i} \log \boldsymbol{o}_{s_i}$ We optimize jointly the high and low resolution augmentation as follows:

$$\mathcal{L}(\theta, \phi) = \sum_{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}} \sum_{\{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_c\}} \mathcal{L}(\boldsymbol{o}_{s_i}, \boldsymbol{o}_{t_i}) \tag{5}$$

We obtained $\boldsymbol{o}_{s_i}$ and $\boldsymbol{o}_{t_i}$ from the Eq-3. The Eq-5 is optimized w.r.t. the student parameter θ i.e.: $\min_\theta \mathcal{L}(\theta, \phi)$ Note that the teacher model is not learned, and it uses exponential moving averaging to update its parameter and it is given as: $\phi_t \leftarrow \lambda\theta_t + (1-\lambda)\phi_t$; where λ is hyperparameter. The cross-entropy loss minimize the oscillate [4], and the model converges faster.

4. Data and Statistics

Product Category	Product Count	Training Images	Test Images
Women's Dress	94,393	534,057	176,993
Shirt	84,775	520,474	157,829
Furniture	56,161	334,589	80,024

Table 1. Datasets and their statistics for the three diverse datasets *Women-dress*, *Shirt* and *Furniture*.

To evaluate the proposed framework at scale, we considered three product categories - shirts, women's dress and furniture of a popular e-commerce service. The statistics are provided in the Table 1, For more details, refer to supplementary sheet.

5. Experiment and Results

In order to evaluate the performance of our model, we conducted rigorous experiments on a highly diverse dataset. The details of implementation, baselines and Evaluation metric are provided in the supplementary.

5.1. Evaluation Metric

To evaluate the models, we selected the attributes with good fill rate and collected the data for these attributes for each product type (task), to train a classification model. We measure the accuracy, Recall@90 (R@90) and Recall@95 (R@95) of the product over the selected attribute individually and report the result. Similarly, we define the evaluation metric to measure the color invariance. To our best knowledge, this is the first metric to measure the color invariant property for the model. We provide the details in the supplementary.

5.2. Results

We have evaluated the learned self-supervised embedding for the two scenario, that are given as follows:

		Pattern						Sleeve					
		Linear			Non-liner			Linear			Non-linear		
		Acc	R@90	R@95	Acc	R@90	R@95	Acc	R@90	R@95	Acc	R@90	R@95
Shirt	BYOL	56.31	04.10	02.24	59.91	6.494	01.70	81.30	71.14	60.15	83.37	73.51	65.46
	SimCLR	78.02	54.88	10.20	78.51	60.15	18.40	81.00	66.47	40.44	83.29	71.89	59.06
	DINO	80.29	**61.21**	25.87	83.19	71.62	50.71	91.39	78.81	77.81	93.93	88.59	81.32
	SkiLL	**81.01**	59.90	**31.07**	**85.66**	**76.76**	**63.52**	**92.81**	**79.36**	**78.75**	**95.14**	**91.99**	**86.93**
Women Dress	BYOL	66.77	14.96	01.24	66.95	15.55	02.43	34.46	00.03	00.03	44.99	00.87	00.34
	SimCLR	81.29	64.08	44.01	82.47	67.10	50.93	50.23	01.71	00.16	53.92	08.20	02.65
	DINO	83.11	67.55	48.28	90.87	87.92	79.96	62.26	17.65	04.08	78.19	58.34	37.09
	SkiLL	**83.99**	**68.04**	**52.30**	**92.32**	**91.34**	**84.46**	**70.97**	**38.32**	**18.67**	**85.57**	**75.71**	**61.21**

Table 2. Results for *Shirt* and *Women-dress* product type, evaluated for 2 attributes: Pattern & Sleeve which has 20 and 6 classes respectively.

	Other Hard-Line (OHL) Subcategories					
	Linear			Non-liner		
	Acc	R@90	R@95	Acc	R@90	R@95
BYOL	56.45	27.46	20.39	62.14	33.44	27.37
SimCLR	72.30	49.88	38.53	75.19	54.61	43.08
DINO	83.10	70.52	58.25	85.04	75.34	64.23
SkiLL	**83.77**	**71.79**	**61.35**	**86.02**	**77.45**	**65.32**

Table 3. Hard-line Subcategories classification result, there are 19 subcategories on the hard-line dataset.

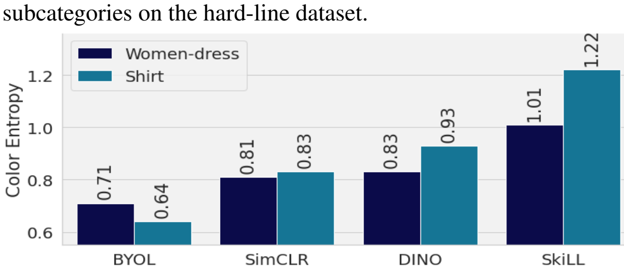

Figure 7. The result for the color invariant measurement, similar in style but agnostic to color shows high color entropy.

5.2.1 Class Discrimination

The results for class discrimination on the Shirt, Women-Dress, and Hard-Line datasets were evaluated using the metric discussed in Section 5.1. We have shown the results for these categories in Table 2 and 3. The CNN network-based models BYOL and SimCLR showed reasonable results for the easy categories of hard-line and shirt sleeves. However, for the pattern and women-dress sleeve categories, the models performance degraded significantly. These models struggled to capture the fine-grain details and discriminate between small variations in the complex categories. However, the proposed model easily capture these information and outperforms the CNN baselines.

5.2.2 Color Invariant

We evaluated the effectiveness of our learned embeddings in terms of color invariance by using the metrics outlined in Section-5.1. The results depicting the average color entropy across all clusters appear in Figure-7. We can see that SkiLL, demonstrates the highest level of color agnosticism compared to the other methods.

6. Ablations

Local Augmentation and Preprocessing: The local augmentation plays a key role in the specific product type. We have conducted the ablation over the proposed task specific augmentation for the challenging dataset women-dress and pattern attribute type. We observe that if from $SkiLL$ we remove the local augmentation ($SkiLL \setminus \mathcal{L}$) the model performance degraded. The $SkiLL$ without pre-processing ($SkiLL \setminus \mathcal{H}$) significantly reduce the model performance. The results for the women dress are shown in the Table-4 for the linear and non-linear evaluation.

Color Agnostic: To learn the color agnostic embedding, we incorporate the *RandomGray*, *ColorJitter* and *GaussianBlur* as a global transformation with the probability $p_1 = 0.5$, $p_2 = 0.8$ and $p_3 = 0.5$ respectively. We observe that for small value of $p_1 = p_2 = p_3 = 0.1$ the color agnostic property of the model significantly degraded and color entropy reduce from 1.01 to 0.90. Therefore, these global augmentations are necessary to achieve the color agnostic embedding.

	Linear			Non-liner		
	Acc	R@90	R@95	Acc	R@90	R@95
$SkiLL \setminus \mathcal{H}$	83.67	**68.73**	50.85	90.92	87.81	78.59
$SkiLL \setminus \mathcal{L}$	83.74	68.15	52.27	91.91	89.25	79.02
SkiLL	**83.99**	68.04	**52.30**	**92.32**	**91.34**	**84.46**

Table 4. Ablation on the women-dress for the pattern classes, without local augmentation ($SkiLL \setminus \mathcal{L}$) model performance degraded.

7. Conclusions

In this paper, we addressed the challenge of learning visual similarity in terms of style and pattern-type, agnostic to other properties, such as color, without the use of labeled data. We proposed a SSL approach using knowledge distillation with the transformer architecture as the backbone for the teacher and student networks. To solve the desired problem we carefully designed an augmentation space containing both local and global augmentations. The local augmentations were task-specific, allowing us to select the appropriate ones for a particular task in combination with the global augmentations. We incorporated the *RandomGray*, *ColorJitter*, and *GaussianBlur* global augmentations to achieve a color agnostic feature space. Our proposed augmentation outperformed the recent baseline by a significant margin. We proposed a novel metric to measure the model's color agnostic property. The ablation shows the importance of the proposed component.

References

[1] Paul Baltescu, Haoyu Chen, Nikil Pancha, Andrew Zhai, Jure Leskovec, and Charles Rosenberg. Itemsage: Learning product embeddings for shopping recommendations at pinterest. *arXiv preprint arXiv:2205.11728*, 2022. 2

[2] Cristian Buciluă, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006. 2

[3] Xuefei Cao, Bor-Chun Chen, and Ser-Nam Lim. Unsupervised deep metric learning via auxiliary rotation loss. 2019. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 1, 2

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 2

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020. 1, 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 2

[10] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[11] Ujjal Kr Dutta, Sandeep Repakula, Maulik Parmar, and Abhinav Ravi. A tale of color variants: Representation and self-supervised learning in fashion e-commerce. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12482–12488, 2022. 2

[12] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 2

[13] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 1, 2

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent a new approach to self-supervised learning. In *Proceedings of the 34th NeurIPS*, 2020. 1, 2

[15] Geonmo Gu and Byungsoo Ko. Symmetrical synthesis for deep metric learning. 01 2020. 2

[16] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[18] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[19] Elad Levi, Tete Xiao, Xiaolong Wang, and Trevor Darrell. Reducing class collapse in metric learning with easy positive sampling. 06 2020. 2

[20] Yang Li, Shichao Kan, and Zhihai He. Unsupervised deep metric learning with transformed attention consistency and contrastive clustering loss. 08 2020. 2

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

[22] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. pages 6397–6406, 06 2020. 2