Simple Unsupervised Knowledge Distillation With Space Similarity

Aditya Singh[1] and Haohan Wang[2]

[1] `aditya.91.singh@gmail.com`
[2] School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, IL, USA
`haohanw@illinois.edu`

Abstract. As per recent studies, Self-supervised learning (SSL) does not readily extend to smaller architectures. One direction to mitigate this shortcoming while simultaneously training a smaller network without labels is to adopt unsupervised knowledge distillation (UKD). Existing UKD approaches handcraft preservation worthy inter/intra sample relationships between the teacher and its student. However, this may overlook/ignore other key relationships present in the mapping of a teacher. In this paper, instead of heuristically constructing preservation worthy relationships between samples, we directly motivate the student to model the teacher's embedding manifold. If the mapped manifold is similar, all inter/intra sample relationships are indirectly conserved. We first demonstrate that prior methods cannot preserve teacher's latent manifold due to their sole reliance on L_2 normalised embedding features. Subsequently, we propose a simple objective to capture the lost information due to normalisation. Our proposed loss component, termed **space similarity**, motivates each dimension of a student's feature space to be similar to the corresponding dimension of its teacher. We perform extensive experiments demonstrating strong performance of our proposed approach on various benchmarks.

Keywords: Unsupervised · Knowledge Distillation · Space Similarity

1 Introduction

In recent years, the development of self-supervised learning (SSL) has allowed networks to be trained on larger datasets without labels, leading to generic representations that are task agnostic and achieve superior downstream performances once fine-tuned [14, 32, 41]. As a result, SSL is an active area of reasearch. For real-time inference, such as in the domain of autonomous driving, industrial automation *etc.* often small sized networks are deployed. However, these networks do not readily benefit from SSL due to their smaller number of parameters, which can hinder their ability to learn underlying discriminative representations effectively [18].

To address this issue, Fang et al. [18] propose an unsupervised knowledge distillation (UKD) framework called SEED that allows smaller networks to take
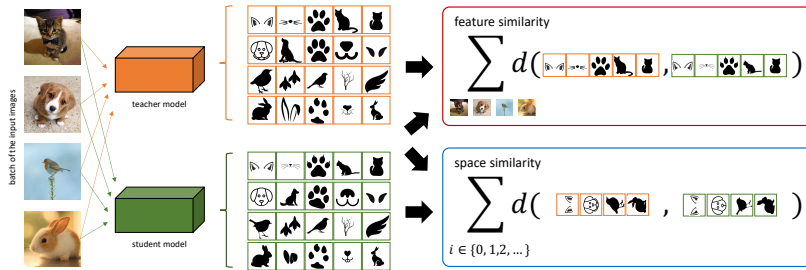
Fig. 1: The proposed CoSS (feature similarity + space similarity) distillation framework. In the graphic, we demonstrate similarity of one pair of corresponding feature dimension being maximised. We perform this maximisation for every corresponding pair for the teacher and student.

advantage of the large amount of data for pre-training. Since the introduction of SEED, many other approaches have followed suite [34, 66]. A common theme amongst many existing UKD methods is that they rely on manually constructing and conserving similarity relationships between training samples. Often this preservation is achieved via. leveraging an embedding queue ($\sim 100k$ in length) which stores features of previously encountered training samples. This conservation of relationships (either at a local or global scale) can be perceived as an attempt to mimic the teacher's manifold. A perfect replication of a teacher's embedding manifold will imply that the pairwise similarity relationships are preserved by the student. Moreover, modelling of the manifold will also preserve those relationships/properties which were otherwise overlooked. However, we observe that, due to the normalization of embedding features, an essential step in existing training recipes, the modelling of teacher's manifold by existing approaches is imperfect (details in Sec. 3).

In this paper, we hypothesize that the knowledge of the teacher is not only in the relationship of the samples but also the manner in which these are mapped onto the latent space *i.e.* the embedding manifold. An alignment of the embedded manifold will imply that the student has learnt to map inputs in the 'same' way as the teacher onto the latent space which will indirectly preserve relationships of interest.

We find that due to sole reliance on L_2 normalized embedding features, existing methods cannot model teacher's embedding manifold. This is because normalization is a non-invertible mapping which eliminates the information and structure held by the original manifold. As a solution, in this paper, we propose a simple **space similarity** objective which works in conjunction with a traditional cosine similarity loss computed on the features. In space similarity, for each feature dimension of a student, we aim to maximise its similarity to the corresponding feature dimension of the teacher. Therefore, space similarity preserves the spatial information, while, the conventional cosine similarity of fea-

tures ensures that the representations learnt are consistent and aligned. Fig. 1 highlights our proposed approach.

Our main contributions are as follows:

– We introduce CoSS, a space similarity inclusive novel objective which motivates the student to mimic its teacher's embedding structure.
– We discuss the limitation imposed by only relying on L_2 normalized features for learning the manifold.
– The simplicity of our approach does not impede the final performance of trained students. We report state of the art results on various UKD benchmarks

The structure of the paper is as follows. In the subsequent Sec. 2, we delve deeper into prior work. In Sec. 3, we motivate the importance of space similarity. In Sec. 4, we provide the details of our simple yet effective method. In Sec. 5, we report results on various benchmarks. We discuss the implications of our work and findings in Sec. 6. Lastly, we conclude with final remarks in Sec. 7.

2   Related Work

2.1   Logit Based Distillation

Early solutions for distillation have been designed for a fully supervised teacher and leverage the output space of the teacher model. Soft-label distillation [27] is amongst the first work towards training smaller networks with guidance from larger teacher networks. Apart from supervised loss from the ground-truth labels, it minimises cross-entropy between the teacher's and the student's output logit distribution. Many methods since then have utilised the output logits to improve the knowledge distillation performance [10, 31, 40, 67, 73].

Huang et al. [29] propose a correlation based formulation which is very similar to ours. The key difference apart from the teacher (supervised vs. unsupervised) is that they normalise the logits (via. softmax) and then compute inter and intra-class similarities. Whereas, we independently normalize spatial and feature dimensions of the embedding features. From the perspective of computing intra-class similarity, it is logical to apply the softmax beforehand for generating class-wise scores, however, when operating on the embedding space, any normalisation on the features alters the space information as well.

2.2   Feature Based Distillation

Feature based distillation approaches leverage the internal representations produced by the teachers and the students.

Many novel approaches have incorporated internal layers, attention mechanisms etc. to match various novel objectives between student and the teacher [1, 24, 30, 33, 35, 55, 58, 68, 72]. However, as most of these approaches require careful selection of an appropriate statistic, it can be a drawback in practice for defining the distillation procedure for newer architectures. Many existing approaches

also utilise local and global relationships [9, 46, 48, 49, 69] in a metric learning framework. They mostly differ in the devised relationships for optimisation. LP [5] further develops the idea of FitNets and introduces locality-preserving loss, which relies on identifying K-nearest neighbours within the training batch. Lastly, there are also methods which utilize self-supervision for knowledge distillation. CRD [57], WKD [6] and SSKD [65] fall into this category. A common theme for all these methods is that they focus on distillation of a teacher trained with supervision and thus often operate in conjunction with a supervised objective. Though they can be utilised for performing distillation of an unsupervised teacher, they have not yet been evaluated in such a setting. A quick evaluation (see supplementary document) of these methods shows that they do not readily adapt to a fully unsupervised setting.

SEED [18] is the first work, to the best of our knowledge, that attempts to distill knowledge in an unsupervised teacher. They perform knowledge distillation of a self-supervised teacher by minimizing the divergence between the similarity response of teacher and student on a common embedding queue. CompRess [34] introduces two feature queues, one each for the teacher and the student. SimReg [42] demonstrates the applicability of a pairwise mean squared error minimisation between the latent embeddings of teacher and student models. Our work leverages and extends the pairwise feature matching objective for enforcing manifold similarity. AttnDistill [61] is a similar non-contrastive UKD method which focuses solely on vision transformers [15]. It leverages the multi-head self-attention mechanism during training. DisCo [20] performs a consistency regularization between augmented versions of the input in addition to unsupervised distillation. BINGO [66] is a two-stage method for performing unsupervised distillation. In the first stage it computes k-nearest neighbours for each training sample in order to construct a bagged dataset. In the second (distillation) stage, it adopts a contrastive distillation approach to minimise divergence between samples from the same bag while increasing distance w.r.t samples from different bags. SMD [39] focuses on mining hard positive and negative pairs rather than operating on all pairs for distillation. To counter the influence of wrongly assigning positive and negative labels, it utilizes a weighting strategy. PCD [28], a recently proposed contrastive method, specializes in dense prediction tasks like image segmentation and object detection.

2.3 Key Differences

Our proposed method is designed for unsupervised distillation, but we believe it stands out from existing methods in both supervised and unsupervised regimes, despite its simplicity. In particular, our approach focuses on directly motivating the student to learn its teacher's latent manifold. As a quantitative summary, CoSS differs from many existing UKD methods in the **absence** of (i) feature queues, (ii) contrastive objectives, (iii) and, heavy augmentations.

3 Motivation

In deep learning, it is generally accepted that neural networks learn a mapping from high dimensional space to a low dimensional manifold. Moreover, assumption of a locally euclidean manifold in unsupervised learning is fundamental but rarely articulated.

For example, many unsupervised learning methods employ manifold learning based data visualisations[44, 74]. These manifold learning approaches assume the embedding manifold to be locally eucldiean[4, 59]. This assumption of the manifold being locally euclidean, allows us to treat the embedded manifold as a topological manifold. Here, we present a simple argument to show that methods solely relying on L_2 normalized features cannot learn the teacher's embedding structure reliably.

Definition 1. *Two topological spaces $\mathcal{X}$, $\mathcal{Y}$ are homeomorphic if there exists a mapping $f : \mathcal{X} \to \mathcal{Y}$ s.t. f is continuous, bijective and its inverse function f^{-1} is also continuous.*

Homeomorphism[51] defines the concept of similarity (equivalence) between the two topological spaces. For methods which only rely on normalized cosine similarity, the student's normalized manifold and the original teacher's manifold are not homeomorphic. This is because the operation of L_2 normalisation is not a homeomorphism. It is not continuous, bijective, and lacks a continuous inverse. A straight forward example to support it is that, after normalization, all points lying on the ray starting from the origin will be mapped onto the same point on a hypersphere. Hence, minimisation of an objective operating on the normalized space will not preserve the original **un-normalized** structure. Existing UKD methods of SEED, DisCo, and BINGO leverage the normalized features during their intermediate steps while SMD, SimReg and AttnDistill perform regression on the normalized feature as an additional training objective. Since, the normalisation erases the original structure, an alternative strategy to retain the spatial information is by preserving similarity along the spatial dimensions.

4 Method

Our motivation centers around the idea of imposing *homeomorphism* between the manifolds of a teacher and student. In order to achieve this, we propose a two step training method. In the first (offline) step, we compute k-nearest neighbours for the training samples. In the second stage, we perform distillation.

4.1 Offline Pre-processing

Typically, during training, a mini-batch is often composed of randomly selected samples without replacement from the training set $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$. The

local structure of the manifold cannot be well established if the local neighbour-hood information is missing.

Hence, in-order to utilise local neighbourhood information during distillation, we append randomly selected k samples (without replacement) for each x_i in a mini-batch from the k-nearest set Ω_i^k.

Computation of Ω_i^k is an offline process and is performed prior to the distillation step. Utilizing the teacher model f_t, we first compute a similarity matrix $\mathbf{S} \in \mathbf{R}^{N \times N}$. Where, $S_{ij} = \hat{f}_t(x_i) \cdot \hat{f}_t(x_j) \ \forall \ \mathcal{X}$ and $\hat{\cdot}$ denotes a L_2 normalized vector. The embeddings, $f_t(x) \in \mathcal{R}^d$, is the response gathered at the penultimate layer of a network (typically after global average pooling [37]). We define the nearest neighbourhood set Ω_i^k as:

$$\Omega_i^k = \arg\max(S_{i\cdot}, k), \tag{1}$$

where $\arg\max(\cdot, k)$ returns the indices of top k items in $S_{i\cdot}$. We shall be using these pre-computed nearest neighbour sets during the distillation process.
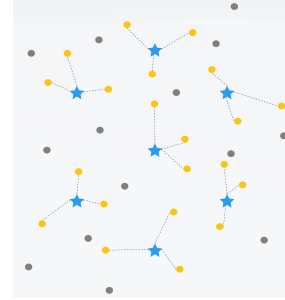
4.2   Training Objectives

Similar to the teacher, we define the student neural network as f_s with the output dimension as d_s. If $d_s \neq d_t$, we can use a small projection head for the student which can then be discarded after the distillation process [18, 20, 66]. We thus replace d_t and d_s by d.

Firstly, for a mini-batch B, we append additional samples to it from $\Omega_i^k \ \forall \ i = 1, 2, \cdots |B|$ and denote the updated mini-batch as $\bar{B}$ (Figure 2 illustrates the enhancement of the training batch). We denote the embedding representations of all inputs in $\bar{B}$ generated by the teacher and student as matrices A_t and $A_s \in R^{bk \times d}$ respectively. Here, $A_\cdot^i$ is the embedding output $f_\cdot(x_i)$. We then compose a matrix of normalized feature vectors $\hat{A}_\cdot = [\hat{A}_\cdot^0, \hat{A}_\cdot^1, \dots \hat{A}_\cdot^{bk}]^T$. The widely known normalized cosine similarity loss on features is defined as:

Fig. 2: The training batch is composed of random samples ($\star$) and their nearest k samples ($\bullet$).

$$\mathcal{L}_{co} = -\frac{1}{bk} \sum_{i=0}^{bk} cosine(\hat{A}_s^i, \ \hat{A}_t^i). \tag{2}$$

The loss computes the cosine similarity between corresponding embeddings of the teacher and student thus, performing a pairwise alignment of features in the normalized embedding manifold. Due to its simplicity and effectiveness $\mathcal{L}_{co}$ is employed by a number of existing UKD methods [39, 42, 61].

To define space similarity, we first construct the transpose matrix of features $Z. = A^T_.$ and its normalized version $\hat{Z}.$. The space similarity loss is

$$\mathcal{L}_{ss} = -\frac{1}{d} \sum_{i=0}^{d} cosine(\hat{Z}_s^i, \ \hat{Z}_t^i). \tag{3}$$

Note, minimizing the loss along the spatial dimension indeed imposes *homeomorphism* as the normalization here scales all data points identically. In case of minimum space alignment loss, $f_s(x_i) = \frac{\alpha}{\beta} f_t(x_i)$ where α and β are scaling (normalization) vectors for the teacher's and student's dimensions respectively. The scaling components for dimensions are not required to be identical, hence, the alignment is similar for each dimension up to a scale. As $\alpha > 0$ and $\beta > 0$, the mapping imposed by $\mathcal{L}_{ss}$ between the corresponding projected data points is continuous, bijective and invertible.

The loss is also very simple to implement and requires one to transpose the feature matrices prior to the normalization. Our final objective is composed of weighted combination of **Co**sine similarity and **S**pace **S**imilarity:

$$\mathcal{L}_{CoSS} = \mathcal{L}_{co} + \lambda \mathcal{L}_{ss} \tag{4}$$

Fig. 1 provides a pictorial representation of our approach. Feature similarity compares the features being extracted for a single sample by the teacher and student whereas, space similarity compares corresponding spatial dimensions.

5 Experiments

5.1 Settings

Dataset: Following SEED, BINGO and others, we report top-1/5 and knn-10 classification accuracy on the ImageNet [13] dataset. For transfer learning, we utilise CIFARs[36], STL-10 [12], Caltech-101 [19], Oxford-IIIT Pets [47], Flowers [43] and DTD [11]. For dense prediction tasks, we employ the PASCAL VOC (`trainval2007`, `trainval2012`, `test2007`) [17] and MS-COCO (`train2017`, `val2017`) [38] datasets. For image retrieval experiments, we employ the Oxford-5k [50] and FORB [63] datasets. We utilize various ImageNet variants namely, ImageNet-v2 [52], ImageNet-Sketch [60] and ImageNet-C [26] to understand robustness of distilled networks.

Teachers: Following SEED, we use the ResNet-50 model pretrained on ImageNet using Moco-v2 [7]. For dense prediction, following PCD, we use the ResNet-50 pretrained on ImageNet using Moco-v3 [8].

Students: We use ResNet-18, ResNet-34, and EfficientNet-B0 as the student architectures. They consist of 10.7M, 20.4M and 4M parameters respectively.

Distillation: We discard the projection head of the teacher and work directly with the 2048 dimensional features from ResNet-50 and add a projection head on top of the students to match the final embedding dimensions. Using

Table 1: Unsupervised distillation of a (self-supervised) ResNet-50 teacher on ImageNet-1K. The *teacher* and *student* correspond to models trained using Moco-v2. Values with ⋄ are replicated results using the corresponding official implementations. We highlight the best student performance in **bold**

Teacher (top-1)					67.40				
Methods	ResNet-18			ResNet-34			Eff-b0		
	T-1	T-5	KNN-10	T-1	T-5	KNN-10	T-1	T-5	KNN-10
Moco-v2	52.20	76.60	36.70	56.80	81.40	41.50	42.20	68.50	30.0
SEED [18]	57.60	81.80	50.12	58.50	82.60	45.20	61.30	82.70	53.11
BINGO [66]	61.40	84.30	**54.16**⋄	63.50	85.70	–	63.74⋄	85.32⋄	54.75⋄
DisCo [20]	60.60	83.70	52.03	62.50	85.40	53.65	66.50	87.60	54.78
SMD⋄ [39]	59.56	83.29	49.69	62.75	85.25	52.61	–	–	–
CoSS (Ours)	**62.35**	**84.81**	53.78	**64.01**	**86.14**	**54.80**	**67.36**	**87.75**	**58.33**

$k=15$, $N=31$, $B=64$, $\lambda=1.0$, an initial learning rate of 0.03 and a cosine scheduler for learning rate decay, we perform distillation for 25 epochs distributed over 4 GPUs. We use the augmentation policy 'mocov2_aug' as described in SEED. We also scale the overall loss by 70.0, as we observed the convergence to be slow.

5.2   Supervised Classification

We evaluate the goodness of a UKD method by performing classification leveraging the embedding features from the trained student network. For the ResNet baselines of BINGO, SEED, and DisCo, we present findings as reported in their corresponding publications. Since BINGO does not provide results for Efficient-Net, we report the reproduced values for BINGO from the official implementation using recommended hyper-parameters. Regarding SMD, their evaluation involves a distinct teacher-student pairing; hence, we employ their official implementation and recommended hyper-parameters to conduct distillation using our selected teacher-student pair. As SimReg is distillation with $\mathcal{L}_{co}$ objective without nearest neighbour sampling, we defer the comparison to the adblation sstudies in Sec. 5.8. We do not compare our work with AttnDistill as it is a solely vision transformer based approach which utilises extended training (up to 800 epochs).

Linear Evaluation: Following the precedent set by SEED, we freeze the backbone and only learn the final linear classification layer. We use a batch size of 256 and an initial learning rates of 10 and 3 for ResNets and EfficientNet respectively. We perform this fine-tuning for 100 epochs with learning rate reduced by a factor of 10 at 60 and 90 epochs.

As per the results in Tab. 1, CoSS achieves significant improvements over the Moco-v2 trained (Student) baseline both for top-1 and top-5 classification accuracies. Compared to the Moco-v2 student, for top-1 accuracy, our method provides an improvement of 10%, 7% and 15% for ResNet-18, ResNet-34 and

Table 2: Transfer learning evaluation of distilled ResNet-18 and Efficientnet-b0. Here, we report the top-1 accuracy.

Method	CIFAR-10		CIFAR-100		STL-10		Caltech-101		Pets		Flowers		DTD	
	ResNet-18	Eff-b0	ResNet-18	Eff-b0	ResNet-18	Eff-b0	ResNet-18	Eff-b0	ResNet-18	Eff-b0	ResNet-18	Eff-b0	ResNet-18	Eff-b0
SEED	85.27	88.85	62.75	69.87	93.99	94.90	80.26	84.98	76.18	78.81	75.10	88.44	67.34	69.79
BINGO	87.67	89.74	66.14	70.25	94.99	94.75	83.84	86.48	79.24	80.80	83.62	90.35	70.00	70.85
DisCo	88.11	91.63	67.50	73.97	**95.04**	**95.78**	84.65	86.44	77.86	81.20	83.69	89.52	69.89	71.91
SMD	86.47	–	64.42	–	94.24	–	80.59	–	74.59	–	78.97	–	69.31	–
CoSS (ours)	**89.23**	**92.72**	**70.11**	**77.17**	94.11	95.41	**86.62**	**90.26**	**79.98**	**83.09**	**85.78**	**93.52**	**70.43**	**74.04**

EfficientNet-b0 respectively. Comparing CoSS to state of the art methods of BINGO and DisCo, we can also observe CoSS' overall improved performance. For top-1, CoSS consistently provides gains over the closest second student. EfficienetNet-b0, while only possessing 16.3% of the parameters, top-1 results are comparable to the teacher with a difference of only 0.04%.

K-NN evaluation: As per prior studies, we report the result of k nearest evaluation [18, 66]. Following SEED, we choose 10 nearest neighbours from the training samples and assign the majority class as the prediction.

As shown in table Tab. 1, CoSS consistently achieves either state of the art or competing performance for k-NN. We believe that the performance gain of CoSS is a result of direct modelling of the teacher's manifold.

5.3 Transfer Learning

We employ the transfer learning benchmark [16] to evaluate the transferability of learned representations. In this evaluation, we keep the student backbone frozen and exclusively train the final classification layer. Ensuring fairness in evaluation, we conduct a thorough hyper-parameter sweep across various learning configurations to pinpoint the optimal set of values for each model across each dataset.

As depicted in Tab. 2, CoSS students exhibit superior transfer learning accuracies across a wide array of datasets for both ResNet-18 and EfficientNet-b0 architectures. It yields the best top-1 accuracies for 6 and 7 datasets for ResNet-18 and Efficient-B0 students respectively. On STL-10, it provides competitive performance to the baselines.

5.4 Dense Predictions

As previously introduced, PCD is a UKD method aimed at learning a student model focused on dense prediction tasks of object detection and image segmentation. For this evaluation, we follow the protocols as outlined by PCD. We utilize the student ResNet-18 distilled from a Moco-v3 ResNet-50 teacher [8]. The hyper-parameters for distillation remain consistent with those mentioned earlier (see Sec. 5.1). We use the Detectron2 [64] package for training and evaluation. The baseline results are directly sourced from PCD.

On VOC2007+2012 [17] we train C4 backbone with Faster RCNN [54] detector whereas for MS-COCO [38], we train C4 with a Mask R-CNN [25]. For

Table 3: Object Detection and Segmentation evaluation of ResNet-18 students. Values for baselines are reported directly from PCD.

Method	ImageNet	VOC 07+12		MS-COCO	
	T-1	AP_{50}	AP	AP^{bbox}	AP^{mask}
Teacher	74.6	83.0	56.7	37.4	32.8
CompRess [34]	63.9	78.4	50.4	31.4	28.4
DisCo	63.5	72.6	40.1	28.2	25.8
BINGO	64.2	77.8	49.3	31.1	28.2
PCD [28]	65.1	**79.4**	52.1	32.2	29.0
CoSS (Ours)	**67.2**	78.7	**52.4**	**33.9**	**30.2**

VOC2007+2012, we set the initial learning rate at 0.1 and perform the training for $24k$ iterations and evaluate the results on `test2007` set. On COCO [38], following PCD, we fine-tune the detector with an initial learning rate of 0.11 for $90k$ iterations. The corresponding scheduling for both the trainings are also referred to as $1\times$ in Detectron2's documentation. For detailed hyper-parameters such as warm up, augmentations, learning rate decay and image resolutions etc. we refer the readers to PCD.

Results in Tab. 3 show that the CoSS student is able to compete with PCD across various settings even though, it is not specifically designed for dense predictions. On the more challenging dataset of COCO, CoSS outperforms PCD. We believe the strong performance of CoSS comes from the overall better knowledge distillation performance. Comparing, top-1 accuracy of the models on ImageNet, we can observe that CoSS provides $\approx 2\%$ gain over PCD.

5.5   Image Retrieval

In this experiment, we investigate the efficacy of knowledge distillation models in the context of image retrieval. Oxford-5k is a widely used dataset for image retrieval. It consists of 55 query images and ≈ 5000 database images categorized into *medium* and *hard* based on the perceived difficulty of the task. FORB is a recently proposed dataset for image retrieval in the flat object setting. It consists of 8 categories with a total of $14,000$ query images and $54,000$ database images. The queries are labelled as *easy*, *medium* and *hard* based on the difficulty level. For Oxford-5k, we report the mean average precision for *medium* and *hard* instances. Whereas, for FORB, we report the overall mean average precision across *easy*, *medium* and *hard* categories. We use the ResNet-18 student distilled from Moco-v2 ResNet-50 for this evaluation.

Table 4 presents the results of the retrieval task. According to these results, CoSS demonstrates improved mAP performance compared to the baseline models. For the few instances where CoSS is not the top performing model, it achieves a second best rating. These results suggest the potential effectiveness of the la-

Table 4: Image retrieval performance on Oxford-5k and FORB, measured in mAP. We highlight the best student performance in **bold**

Methods	Oxford-5k		FORB							
	Medium	Hard	Animated	Photorealistic	Bookcovers	Paintings	Currency	Logos	Packaged	Posters
SEED	17.19	3.71	0.040	0.184	0.099	0.267	0.108	0.007	0.071	0.100
SMD	16.77	3.53	0.043	0.133	0.093	0.214	0.121	0.006	0.065	0.102
DisCo	18.63	4.12	0.120	0.283	0.176	0.336	0.215	**0.016**	0.129	0.160
BINGO	19.80	**5.84**	0.158	0.299	0.193	0.305	0.243	0.012	0.144	0.172
CoSS	**20.59**	5.50	**0.179**	**0.302**	**0.220**	**0.353**	**0.286**	0.013	**0.155**	**0.205**

Table 5: Robustness evaluation of Moco-v2 ResNet-50 distilled ResNet18s. We report the Top-1 classification accuracy. We highlight the best performance in **bold**.

Method	ImageNet-v2			ImageNet-S	ImageNet-C														
	MF	Tr	Top		brightness	contrast	defocus	elastic	fog	frost	gaussian	glass	impulse	jpeg	motion	pixelate	shot	snow	zoom
SEED	45.46	53.24	60.25	13.36	53.84	48.33	36.75	48.64	36.65	39.74	39.85	33.86	23.86	41.38	39.32	47.14	37.81	30.97	26.67
BINGO	46.80	55.58	62.21	14.49	55.51	49.91	37.45	49.97	37.96	41.51	40.17	33.85	23.70	44.10	39.18	49.39	38.77	32.20	27.33
DisCo	47.66	55.96	62.74	14.40	56.07	50.79	39.05	50.56	38.18	42.32	42.39	36.34	27.24	44.83	41.84	50.10	41.12	33.27	28.43
SMD	47.13	55.40	61.84	13.15	55.51	50.18	38.72	50.50	38.12	41.12	40.44	34.66	24.60	44.22	40.66	49.10	38.92	32.39	28.60
CoSS (ours)	**49.70**	**58.21**	**64.84**	**15.90**	**58.78**	**53.26**	**42.44**	**53.49**	**42.09**	**44.72**	**45.32**	**38.74**	**27.74**	**48.82**	**44.42**	**53.42**	**43.54**	**35.58**	**31.79**

tent representations produced by the student across various scenarios, indicating promising capabilities for real-world applications.

5.6   Out-of-Distribution Robustness

Various studies have indicated that deep learning systems often break down when encountered with data outside the training distribution[21, 26, 60]. Due to the wide spread applicability of deep learning systems in the real world, it becomes important to ensure a high degree of robustness of such systems. In this experiment we explore the robustness of trained students under different kinds of shifts in the input data. ImageNet-v2 is a natural image dataset which closely resembles the sampling process of the original ImageNet dataset. It has been employed by various studies to understand the robustness of models under natural data shift [45, 56]. ImageNet-S consists of sketch images for ImageNet classes. It is a stronger deviation from the types of images present in the training set. ImageNet-C is a synthetically generated dataset for evaluating a model's performance under various forms of corruptions. We utilise corruption=1 for our evaluation.

Results reported in Tab. 5 demonstrate that CoSS is robust across various input distribution shifts. It even outperforms BINGO which employs stronger data augmentations in the form of CutMix [71]. Strong data augmentations such as CutMix have been shown to improve model robustness.

5.7   Qualitative Evaluation

In this experiment, we qualitatively demonstrate the lack of manifold modelling capability of SEED, DisCo and SimReg($\mathcal{L}_{co}$) with a set of toy experiments. We
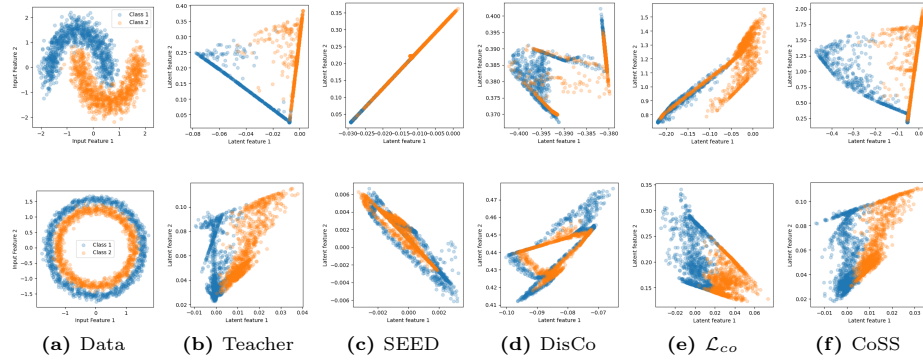
Fig. 3: Plots comparing the latent space of the teacher and different students. Visually, we can assess that though SEED is able to separate the input samples adequately, the learnt mapping is not faithful to the teacher. Whereas, adding the space similarity objective to the standard cosine similarity allows the student to learn a mapping which aligns better with its teacher.

first train teacher models on toy datasets (Two-moons and Circles [2]) using a contrastive objective with positives and negatives determined by the class labels. For the architecture, we consider an MLP with 4 hidden layers, $f : R^2 \rightarrow R^2$ with ReLU activations. Hidden layers perform the following mappings $R^2 \rightarrow R^4 \rightarrow R^8 \rightarrow R^4 \rightarrow R^2$. For the training data, we generate 2500 samples with a noise factor of 0.125. The teacher is trained using constrastive loss for 300 epochs with an initial learning rate of 0.001 using an ADAM optimiser and cosine learning-rate decay. The students are trained using the same hyper-parameters. Using objectives of different students, we perform the unsupervised knowledge distillation. For DisCo, we simulate transformations to generate multiple views of the input by adding perturbations randomly sampled from $\mathcal{N}(0, 0.01)$.

In Fig. 3, we visualize the training data along with the latent spaces of teacher and student models. In the teacher's latent space we can observe few stand out structures. It can also be observed that SEED fails to adequately replicate teacher's manifold. The latent space of SEED and DisCo lacks various details in comparison to the teacher. Pairwise cosine similarity, similar to SEED and DisCo, appears to be ignoring the structure of the teacher's manifold. However, the best quantitative results are observed with cosine similarity in conjunction with space similarity. As mentioned earlier, the embedding manifold is similar up to a scale.

5.8   Ablation Study

Importance of Loss Components: We compare the results of distillation with individual components with the combined objective of CoSS. We report results with distillation of a Moco-v2 ResNet-50 to a ResNet-18 and EfficientNet-b0 student. The hyper-parameters are those provided in Section 5.2. Results in

Table 6: Examining individual loss components

Arch.	$\mathcal{L}_{co}$	$\mathcal{L}_{ss}$	$\mathcal{L}_{coss}$
R-50/R-18	61.85	61.79	**62.35**
R50/Eff-b0	66.41	66.76	**67.36**

Table 7: Importance of nearest sampling

$\mathcal{L}_{CoSS}^{k=0}$	$\mathcal{L}_{CoSS}^{k=15}$
61.89	**62.35**

Table 8: Importance of λ

$\lambda = 0$	$\lambda = 0.5$	$\lambda = 1.0$
61.85	**62.41**	62.35

Table 9: Comparison to soft-label distillation

	ResNet-18		Eff-b0	
	Top-1	KNN-10	Top-1	KNN-10
SLD	59.88	52.67	61.49	56.10
CoSS	**62.35**	**53.78**	**67.36**	**58.33**

Table 10: ResNet-101 teacher distillation

SEED	DisCo	CoSS
58.90	62.30	**63.74**

Tab. 6 highlight the effectiveness of combing both the components over a single loss objective. When using the entire dataset, the degree of improvement differs for different teacher-student architectures. In the supplementary document we share results on CIFAR-100 dataset.

Importance of sampling: We evaluate the impact of nearest neighbour sampling on the training of the student. Utilising ResNet-50/ResNet-18 teacher student pair we perform distillation as per the settings detailed in Section 5.1. To recall, k controls the number of nearest neighbours added to a batch of randomly sampled training images. In this experiment, we compare $k = 0$ and $k = 15$. $k = 0$ implies that the model will be focusing more on the high level structural similarity of the manifold and may miss the finer details. We utilise 100% of the training set for distillation. As results in Tab. 7 indicate, using neighbour hood sampling gives a boost in student's performance. This result highlights the importance of capturing local information when modeling the teacher's manifold in addition to its global structure.

Value of λ: λ controls the contribution of space similarity for the distillation training. We have used $\lambda = 1$ for our main experiments. Here, we report results with $\lambda = \{0, 0.5, 1.0\}$. $\lambda = 0$ is equivalent to only applying $\mathcal{L}_{co}$ or SimReg [42]. We use ResNet-50/ResNet-18 teacher student pair distilled as per settings outlined in Section 5.1. Table 8 indicates that weighing the space component equally to conventional feature similarity gives improved results. Also, the performance remains consistent for $\lambda = \{0.5, 1.0\}$. In the supplementary, we share results with CIFAR-100.

Soft-label Distillation: In this experiment, we compare CoSS with soft-label distillation (SLD) adapted for unsupervised knowledge distillation. To recall, SLD [70] minimises the KL divergence between a student and teacher's output probabilities in a supervised setting. DINO [3] extends this objective for self-supervised learning. Here, we employ the same objective as that of DINO and compare it against CoSS. For SLD, the training hyper-parameters are identical to those applied for SEED and for CoSS the hyper-parameters are provided in 5.1. We use a ResNet-50 teacher and a ResNet-18 student for this experiment.

As shown in Tab. 9, our formulation provides significant performance boost over the SLD baseline.

ResNet-101 teacher: Using a ResNet-101 trained using Moco-v2 as the teacher, we perform the distillation onto a student ResNet-18. We follow the distillation hyper-parameters as outlined in Sec. 5. For baselines, we report the values as sourced from DisCo. The findings in the Tab. 10 demonstrate that our approach readily extends to a different teacher architecture.

6   Discussion

In this work, we address the challenging and critical problem of unsupervised knowledge distillation. While prior studies have predominantly focused on establishing distillation-worthy relationships among samples, our approach takes a distinct perspective by directly modelling the teacher's manifold. By doing so, we aim to indirectly preserve relationships between the samples.

Central to our approach is the utilization of space similarity to establish a form of *homeomorphism* between the projections of the student and teacher models. This homeomorphism ensures that the fundamental topological properties of the learned representations are preserved during the distillation process which otherwise are lost due to sole reliance on L_2 normalized embedding features. Homeomorphism, however, only ensures that the alignment of individual spatial dimensions is upto a scale. Moving forward, we aim to explore even stronger constraints on learning student topologies, which could potentially improve the effectiveness and robustness of our approach.

Our experimental results demonstrate the effectiveness of CoSS in training student models that not only perform well on the training distribution but also deliver competitive performance on various downstream tasks. For instance, CoSS outperforms PCD on the COCO dataset for dense prediction tasks. Furthermore, we believe that integrating CoSS into existing frameworks, such as PCD, could further enhance the outcomes of dense prediction tasks, opening up new avenues for research and application.

In this study, our primary focus has been on the domain of computer vision. However, the rapid development of unsupervised large models in natural language processing presents an intriguing opportunity [22, 23, 53, 62]. Exploring the potential transference of our proposed method to this domain is a direction we leave for future exploration.

7   Conclusion

In this paper, we are inspired by the necessity to distill large models trained with self-supervised learning into smaller models. However, since the labels used to train these large models are typically unavailable, we investigate knowledge distillation in a purely unsupervised setting. In this setting, we demonstrate that unsupervised feature distillation can be achieved without the need to store a feature queue, and directly modelling the teacher's manifold.

Bibliography

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *CVPR*, pages 9163–9171, 2019.

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

[4] Lawrence Cayton et al. *Algorithms for manifold learning*. eScholarship, University of California, 2008.

[5] Hanting Chen, Yunhe Wang, Chang Xu, Chao Xu, and Dacheng Tao. Learning student networks via feature embedding. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):25–35, 2021. `https://doi.org/10.1109/TNNLS.2020.2970494`.

[6] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. Wasserstein contrastive representation distillation. In *CVPR*, June 2021.

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

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9620–9629, 2021. URL `https://api.semanticscholar.org/CorpusID:233024948`.

[9] Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[10] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[11] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[12] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single Layer Networks in Unsupervised Feature Learning. In *AISTATS*, 2011.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[14] Carl Doersch, Abhinav Kumar Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *ICCV*, 2015.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[16] Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How Well Do Self-Supervised Models Transfer? In *CVPR*, 2021.

[17] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, January 2015.

[18] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *ICLR*, 2021.

[19] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2006.

[20] Yuting Gao, Jia-Xin Zhuang, Shaohui Lin, Hao Cheng, Xing Sun, Ke Li, and Chunhua Shen. Disco: Remedy self-supervised learning on lightweight models with distilled contrastive learning. *European Conference on Computer Vision(ECCV)*, 2022.

[21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`.

[22] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=5h0qf7IBZZ`.

[23] Shivanshu Gupta, Yoshitomo Matsubara, Ankit Chadha, and Alessandro Moschitti. Cross-lingual knowledge distillation for answer sentence selection in low-resource languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, July 2023.

[24] Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=8qePPvL1VY`.

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE Computer Society, 2017. ISBN 978-1-5386-1032-9. URL `http://dblp.uni-trier.de/db/conf/iccv/iccv2017.html#HeGDG17`.

[26] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[28] Junqiang Huang and Zichao Guo. Pixel-wise contrastive distillation. In *ICCV*, pages 16359–16369, 2023.

[29] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

[30] Zehao Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[31] Ying Jin, Jiaqi Wang, and Dahua Lin. Multi-level logit distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24276–24285, 2023.

[32] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2021.

[33] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, pages 2760–2769, 2018.

[34] Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. In *NeurIPS*, 2020.

[35] Animesh Koratana, Daniel Kang, Peter Bailis, and Matei Zaharia. LIT: Learned intermediate representation training for model compression. In *ICML*, pages 3509–3518, 09–15 Jun 2019.

[36] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[37] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *ICLR*, 2014.

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[39] Hao Liu and Mang Ye. Improving self-supervised lightweight model learning via hard-aware metric distillation. In *ECCV*, pages 295–311. Springer, 2022.

[40] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, pages 5191–5198, 2020.

[41] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.

[42] K L Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation. In *British Machine Vision Conference (BMVC)*, 2021.

[43] Maria-Elena Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *CVPR*, volume 2, pages 1447–1454, 2006.

[44] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[45] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019.

[46] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pages 3967–3976, 2019.

[47] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[48] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, pages 268–284, 2018.

[49] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. Correlation congruence for knowledge distillation. In *ICCV*, pages 5007–5016, 2019.

[50] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[51] Henri Poincaré. *Papers on topology: analysis situs and its five supplements*, volume 37. American Mathematical Soc., 2010.

[52] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019.

[53] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020.

[54] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *NeurIPS*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[55] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[56] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

[57] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *ICLR*, 2020.

[58] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *ICCV*, pages 1365–1374, 2019.

[59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[60] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019.

[61] Kai Wang, Fei Yang, and Joost van de Weijer. Attention distillation: self-supervised vision transformer students need more guidance. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 666. BMVA Press, 2022. URL `https://bmvc2022.mpi-inf.mpg.de/666/`.

[62] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, and Han Hu. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21970–21980, October 2023.

[63] Pengxiang Wu, Siman Wang, Kevin S Dela Rosa, and Derek Hao Hu. FORB: A flat object retrieval benchmark for universal image embedding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL `https://openreview.net/forum?id=c3kuX7ltzr`.

[64] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[65] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *ECCV*, 2020.

[66] Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. In *International Conference on Learning Representations*, 2022.

[67] Chenglin Yang, Lingxi Xie, Chi Su, and Alan L Yuille. Snapshot distillation: Teacher-student optimization in one generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2859–2868, 2019.

[68] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *CVPR*, pages 4133–4141, 2017.

[69] Lu Yu, Vacit Oguz Yazici, Xialei Liu, Joost van de Weijer, Yongmei Cheng, and Arnau Ramisa. Cvpr. In *CVPR*, 2019.

[70] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 2020.

[71] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[72] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

[73] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4320–4328, 2018.

[74] Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4912–4921, 2021.