Exploring the Limits of Deep Image Clustering using Pretrained Models

Nikolas Adaloglou*
adaloglo@hhu.de

Felix Michels*
felix.michels@hhu.de

Hamza Kalisch
hakal104@hhu.de

Markus Kollmann
markus.kollmann@hhu.de

Heinrich Heine University
Düsseldorf, Germany

Abstract

We present a general methodology that learns to classify images without labels by leveraging pretrained feature extractors. Our approach involves self-distillation training of clustering heads, based on the fact that nearest neighbours in the pretrained feature space are likely to share the same label. We propose a novel objective that learns associations between image features by introducing a variant of pointwise mutual information together with instance weighting. We demonstrate that the proposed objective is able to attenuate the effect of false positive pairs while efficiently exploiting the structure in the pretrained feature space. As a result, we improve the clustering accuracy over k-means on 17 different pretrained models by 6.1% and 12.2% on ImageNet and CIFAR100, respectively. Finally, using self-supervised vision transformers we achieve a clustering accuracy of 61.6% on ImageNet. The code is available at https://github.com/HHU-MMBS/TEMI-official-BMVC2023.

1 Introduction

Given a plethora of publicly available pretrained vision models, we ask the following questions: a) how well-structured is the feature space of pretrained architectures with respect to label-related information, and b) how to best adapt this structure to unsupervised tasks. To answer these questions, we focus on unsupervised image classification, also known as image clustering. Image clustering is the task of assigning a semantic label to an image, given an a priori finite set of classes. Ultimately, image clustering consists of simultaneously learning the relevant representations and the cluster assignments.

To begin addressing the aforementioned questions, we present the key challenges regarding image clustering. First, given that we can roughly estimate the number of ground-truth labels, the underlying distribution among classes is hard to infer from the data, which is typically assumed to be uniform. Second, the images should ideally be classified consistently (images of the same class are grouped together) and confidently (one-hot prediction probability).
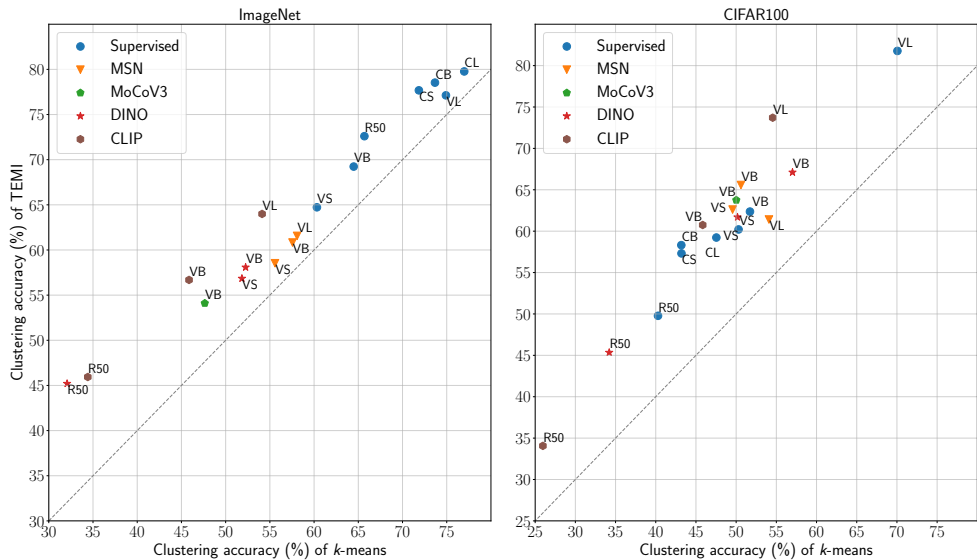
*Equal contribution.

Figure 1: **Clustering accuracies on ImageNet (left) and CIFAR100 (right) across 17 pretrained models.** Supervised and self-supervised models (MSN, MoCoV3, DINO) were pretrained on ImageNet. R50 stands for ResNet50 [17], C for ConvNext [26], and V for Vision Transformer [13]. Small (S), Base (B), and Large (L) indicate the size of the models. The vertical distance of each data point to the diagonal (dashed line) shows the improvement of our method (TEMI) over k-means. Best viewed in color.

Consistency can be achieved by either learning features that are invariant under transformations of the same image (e.g. cropping, colour jitter, etc.), or invariant w.r.t. to substitution by other images that belong to the same semantic class. Consequently, clustering methods are generally prone to degenerate solutions [1]. In other words, samples tend to collapse into a single cluster or the prediction probability spreads out uniformly.

It is well-established that representation learning plays a critical role in image clustering [6], which is achieved with self-supervised learning [7, 16, 18, 39]. Recent studies have demonstrated that self-supervised features are typically more transferable to new tasks than features from supervised learning [15]. The frequently used joint-embedding architectures [5, 16] are by design invariant to strong image transformations that preserve label-related information. That renders these architectures as promising candidates for image clustering [34], which has not been thoroughly explored at scale [40]. Even though self-supervision [15] and vision transformers (ViTs) [29] have been separately established for representation learning, limited research has been conducted to study self-supervised ViTs or vision-language models (i.e. CLIP [32]).

How to adapt a pretrained model for image clustering is non-trivial. For instance, it is well known that k-means is sub-optimal, as it often leads to imbalanced clusters [34] since it is primarily suitable for evenly scattered data samples around their centroids [37]. On the other hand, deep image clustering methods normally rely on pairs by mining the nearest neighbours (NN) based on their feature similarity [14, 20]. Still, images that are close in the feature space do not always share the same semantic class [34] and therefore must be considered as noisy pairs.

In this paper, a two-stage method that extends the existing multi-stage clustering approaches is proposed. In contrast to [34], where features are learned from scratch for each dataset, we show that multi-stage clustering approaches can leverage pretrained models trained on larger-scale datasets (Fig. 1) and focus on learning the cluster assignments. To this end, a self-distillation clustering framework is introduced using a novel objective based on pointwise mutual information and instance weighting. Second, a comprehensive experimental study across models and datasets is conducted. Therein, we report an average gain of 6.1% and 12.2% in clustering accuracy compared to k-means on ImageNet and CIFAR100 across 17 pretrained models, as illustrated in Figure 1. Overall, we show that ViTs capture the most transferable label-related features. Finally, we find that self-supervised ViTs [2] achieve state-of-the-art results (61.6% clustering accuracy) on ImageNet, without using the ground-truth labels or external data.

2 Related Work

Single-stage Deep Image Clustering Methods. Deep image clustering approaches can be roughly divided into single and multi-stage methods. The majority of single-stage methods alternate between learning the features and the clusters, i.e. in an expectation-maximization manner. For instance, in DAC, [6] formulate a binary pairwise-classification task, where at each iteration pairs are selected based on their feature similarity. Next, the computed pairs are used to train a convolutional neural network (CNN). In the same direction, in DeepCluster [4], the authors alternate between clustering the features of a CNN with k-means [27] and using the obtained cluster assignments as pseudo-labels to optimize the parameters of the CNN.

Later on, in [38], the authors demonstrate that DeepCluster is prone to degenerate solutions that are avoided via particular hyperparameter choices. To that end, the authors design a multi-step pseudo-label extraction framework, called SeLa. The latter iteratively estimates the pseudo-label assignment matrix under the equipartition constraint. In PCL [25], the authors formulate clustering as learning the cluster centroids with k-means in parallel with optimizing the network via contrastive learning [7]. To overcome the class collision of the negative pairs, [21] extend PCL in a proximal framework called ProPos. ProPos only maximizes the distance between the cluster centroids with contrastive learning, while mining NN for neighbouring sample alignment [16]. However, most of the existing approaches still rely on k-means for estimating the clusters (pseudo-labels).

Several single-stage approaches exist, which aim to jointly learn the feature representations and clusters. Single-step methods are known to be sensitive to weight initialization [11]. In this direction, DCMM is developed [35] to progressively mine NN in the feature space as well as high-confident samples. Another single-stage end-to-end example is IIC, wherein [22] derives a mutual information-based objective for paired data to train a CNN. Nevertheless, the aforementioned approaches only consider stochastic transforms of the same image to obtain a pair. They are hence limited to solely learning invariances w.r.t. image augmentations, which cannot cover the variability of a given class [14]. More recently, [0] presented a single-stage end-to-end method, called SSCN, that employs a variant of the cross-entropy loss.

Multi-stage Deep Image Clustering Methods. Multi-stage methods initially design a pretext task in order to learn semantically meaningful features, such as denoising autoencoders [36]. A major breakthrough in deep image clustering is established by the adoption of contrastive learning [7, 18]. For instance, [34] decouples image clustering into three distinct steps, starting with contrastive learning. Subsequently, the authors train a head to cluster the mined

NN from the extracted features. Lastly, they use the pseudo-labels from the confidently assigned samples to fine-tune the whole architecture. A similar approach, called NNM [41], aims to mine NN from the batch and dataset features.

Recently, [40] leverage self-supervised pretrained ViTs [6] and train a clustering head, which is closer to our method. Nevertheless, their approach (TSP) heavily relies on k-means for the weight initialization phase. Surprisingly, very few image clustering approaches [1, 54, 58] have been successfully applied on ImageNet [12]. Besides, most methods report results only with Resnet50 [17], while superior architectures for image recognition remain unexplored [13, 26].

3 Proposed Method

3.1 Classification Model

Our aim is to learn a probabilistic classifier from pairs of examples that share label-related information. We assume that the data distribution, $p(x)$, is the result of a generative process, $c \sim p(c)$ and $x \sim p(x|c)$, with $p(c)$ the prior probability that an example belongs to a class $c \in \{1,..,C\}$. Consequently, the joint distribution, $p(x,x')$ that a pair of examples, (x,x'), belongs to the same class is given by $p(x,x') = \sum_{c=1}^{C} p(x|c)p(x'|c)p(c)$. We introduce a parametrized probabilistic classifier, $q(c|x)$, that distributes examples $x \sim p(x)$ among classes, with class occupancy given by $q(c) = \mathbb{E}_{x \sim p(x)}[q(c|x)]$. Using Bayes' theorem $q(x|c) = q(c|x)p(x)/q(c)$, the joint distribution, $p(x,x')$, can be approximated by

$$q(x,x') = \sum_{c=1}^{C} q(x|c)q(x'|c)q(c). \tag{1}$$

To estimate the association between x and x' we introduce the pointwise mutual information, $\text{pmi}(x,x')$ [8], defined by

$$\text{pmi}(x,x') := \log \frac{q(x,x')}{p(x)p(x')} = \log \sum_{c=1}^{C} \frac{q(c|x)q(c|x')}{q(c)}. \tag{2}$$

Theorem 1. *If (i) each example $x \sim p(x)$ belongs to one and only one cluster under the generative model $p(x) = \sum_c p(x|c)p(c)$, (ii) the joint distribution $p(x,x')$ is known, and (iii) $q^*(c|x)$ is a probabilistic classifier defined by*

$$q^*(c|x) = \arg\max_{q(c|x)} \mathbb{E}_{x,x' \sim p(x,x')}[\text{pmi}(x,x')], \tag{3}$$

then $q^*(c|x)$ is equal to the optimal probabilistic classifier, $p(c|x) = p(x|c)p(c)/p(x)$, up to a permutation of cluster indices.

The proof can be found in the supplementary material. Theorem 1 states that under condition (i) the knowledge of pairs of examples belonging to the same class suffices to establish an objective for an optimal classification model.

3.2 Self-distillation Clustering Framework

The starting point is a pretrained feature extractor (backbone) $g(\cdot)$ that assigns each example x in the dataset D a feature vector $g(x)$. We mine the k nearest neighbours (k-NN) of x in the

feature space by computing the cosine similarity between $g(x)$ and the feature vectors of all other images in D. We denote the set of k-NN for x by S_x. During training, we randomly sample x from D along with x' from S_x, to generate image pairs that share label information with high probability.

We introduce two clustering heads, a *student head*, $h_s(\cdot)$, and a *teacher head*, $h_t(\cdot)$, that share the same architecture but differ w.r.t. their parameters, θ_s, and θ_t. Each head consists of a three-layer fully connected feed-forward network. The image pairs x, x' are fed to the shared backbone and subsequently in the two heads, $h_s(g(x))$ and $h_t(g(x'))$. The head outputs are converted to probabilistic classifiers, $q_s(c|x)$ and $q_t(c|x')$, using a temperature-scaled softmax function, which for the student's head is given by

$$q_s(c|x) = \frac{\exp(h_s(g(x))_c / \tau)}{\sum_{c'} \exp(h_s(g(x))_{c'} / \tau)}, \tag{4}$$

where τ is the temperature hyperparameter. Unlike previous self-distillation frameworks [5], we use the same temperature $\tau = 0.1$ for both heads. We approximate the pointwise mutual information by

$$\widetilde{\text{pmi}}(x, x') := \log \sum_{c=1}^{C} \frac{q_s(c|x)q_t(c|x')}{\tilde{q}_t(c)}. \tag{5}$$

and estimate $q(c)$ by an exponential moving average (EMA) over batches using the teacher's head

$$\tilde{q}_t(c) \leftarrow m\tilde{q}_t(c) + (1-m)\frac{1}{B}\sum_{i=1}^{B} q_t(c|x_i), \tag{6}$$

with B the batch size and $m \in (0,1)$ a momentum parameter. In practice, we symmetrize Equation (5) to compute the loss function

$$\mathcal{L}(x, x') := -\frac{1}{2}\left(\widetilde{\text{pmi}}(x, x') + \widetilde{\text{pmi}}(x', x)\right). \tag{7}$$

Note that only the parameters θ_s of the student's head are updated using backpropagation. The parameters of the teacher's head, θ_t, are updated by an exponential moving average for the student parameters, θ_s, over past update steps [5, 16]. As a result, $p_t(c|x)$ represents a sufficiently stable target distribution for the student head. In contrast to other self-distillation frameworks [5] no complicated adaptation of softmax temperatures over training is required. Following previous work [54], we employ an ensemble of H independent clustering heads in training (fig. 3), which alleviates the variability stemming from random initialization. For the evaluation, we use the teacher head with the lowest training loss.

3.3 Balancing class utilization

For a dataset D that has been generated using balanced classes, $p(c) = const$, we expect that $\tilde{q}(c) \approx const$, as a consequence of the optimization process. However, in practice, we observe that classes are typically far from uniformly utilized. We suspect that our self-distillation learning framework leads to over-confident class predictions for a fraction of classes in the early training phase.

To limit the effect of aligning the cluster predictions and take into account the individual cluster probability of the mini-batches $\tilde{q}_t^i(c)$, we introduce a hyperparameter $\beta \in (0.5, 1]$ in

Equation (5) to avoid collapsing all sample pairs in a single cluster. Without affecting the optimal solution, we rewrite Equation (5) as

$$\widetilde{\text{pmi}}^i(x,x') = \log \sum_{c=1}^{C} \frac{\left(q_s^i(c|x)q_t^i(c|x')\right)^\beta}{\tilde{q}_t^i(c)}, \tag{8}$$

where $i \in \{1,\dots,H\}$ is the head index.

3.3.1 Intuition for β.

We now provide an explanation of how the above equation addresses the discussed challenges of image clustering. Equation (8) consists of two parts inside the log sum: the numerator $\left(q_s^i(c|x)q_t^i(c|x')\right)^\beta$ encourages the class assignment of a positive pair to align (consistency) and is maximal when this assignment is one-hot. The denominator $\tilde{q}_t^i(c)$ promotes a uniform cluster distribution by up-weighing the summand corresponding to classes with low probability ($\tilde{q}_t^i(c)$ is low). In effect, β balances these two effects by reducing the influence of the numerator, and thus degenerative solutions are avoided.

Note that for $\beta = 0.5$ the loss corresponds to the Bhattacharyya distance [4] if $\tilde{q}_t^i(c) = const$. The Bhattacharyya distance can be minimal even if q_t^i is far from one-hot. Moreover, if utilization of all classes is not required – for example as in the case of overclustering – we set $\beta = 1$. Crucially, β is bounded, and we empirically found that the value of 0.6 consistently avoids degenerative solutions across the majority of datasets, as opposed to existing clustering methods [1, 34]. In additon, we propose an experimental strategy to choose β without access to the ground-truth labels, as explained in Section 4.2. The symmetrized loss from Equation (8) is defined as $\mathcal{L}^i(x,x')$ in analogy to Equation (7).

3.4 Teacher-guided Instance Weighting

As discussed in Section 1, the mined k-NN in the feature space of $g(\cdot)$ tend to be noisy. For this reason, we introduce an instance weighting term for each head i given by

$$w_i(x,x') = \sum_{c=1}^{C} q_t^i(c|x)q_t^i(c|x'). \tag{9}$$

Intuitively, $w_i(x,x')$ acts as a guidance term that assigns a higher weight to true positive pairs compared to false positive ones. Importantly, $w_i(x,x')$ relies only on the predictions of the teacher. The rationale behind this is that model averaging over training iterations tends to produce more accurate predictions [31, 33]. We call this setup teacher-weighted pointwise mutual information (WPMI). The final objective for each separate head i is given by $\mathcal{L}_{\text{WPMI}}^i(x,x') := w_i(x,x')\mathcal{L}^i(x,x')$.

3.4.1 TEMI: Teacher Ensemble-weighted pointwise Mutual Information

To compensate for the noisy NN pairs based on the feature space of g, we further propose to aggregate information from multiple heads, in contrast to previous works [34] that employ independent heads. For this purpose, we compute a scalar for each image pair using the mean

weight across the heads, which is conceptually similar to model ensembling. We thus call this loss TEMI (teacher ensemble-weighted pointwise mutual information) defined by

$$\mathcal{L}^i_{\text{TEMI}}(x,x') := \frac{1}{H}\sum_{j=1}^{H} w_j(x,x')\mathcal{L}^i(x,x'). \tag{10}$$

4 Experimental evaluation

4.1 Datasets, Metrics and Implementation Details

The proposed method (TEMI) is evaluated on five common benchmark datasets, namely CIFAR10, CIFAR20, CIFAR100 [23], STL10 [9], and ImageNet [12]. CIFAR10, CIFAR20 and CIFAR100 contain $50K$ training images, STL10 contains $5K$ training samples, and ImageNet has $1,281,167$ training samples. CIFAR20 has the same training data as CIFAR100 with 20 superclasses derived from the ground-truth labels. We resize all images to 224×224. The training set is used during the optimization phase, while the evaluations are carried out on the validation set. Additional information can be found in the supplementary material.

To quantify the clustering performance, we report the clustering accuracy (ACC), and the adjusted random index (ARI). To estimate the accuracy, the one-to-one mapping between cluster predictions and ground-truth labels is computed by the Hungarian matching algorithm [24]. For our overclustering experiments, we report the adjusted mutual information (AMI). Finally, we establish two baselines: a) k-means and b) the SCAN clustering loss within our self-distillation framework. For a fair comparison with existing methods, we assume to know in advance the number of ground-truth labels. Concerning the hyperparameters, we set $H = 50$ and $\beta = 0.6$ for clustering, while we set $\beta = 1$ for overclustering. We use 25-NN on ImageNet, and 50-NN for the remaining datasets. We used the AdamW optimizer [28] for 200 epochs with a batch size of 512, with a learning rate of 10^{-4}, weight decay of 10^{-4} and report results at the end of training. Unlike previous methods [34], we found that augmentations (RandAugment [10], and the ones from [7]) were not improving the clustering metrics when training with k-NN pairs. Hence, we precomputed the feature representations, which enables training the clustering heads on a single GPU with 12GB of VRAM, within 24 hours.

4.2 Experimental Results

We first present a strategy to choose $\beta \in (0.5, 1]$. As depicted in Figure 2, an accurate model, $p_t(c|x)$ should be able to maintain a high entropy $H(q_t(c))$, while maintaining its discriminative power. To quantify the latter we use the conditional entropy $H(q_t(c|x))$. The lower the value of $H(q_t(c|x))$ the more discriminative the predictions. The extreme case $H(q_t(c|x)) = 0$, corresponds to a one-hot distribution. Thus, we propose to pick the lowest value of β such that $H(q_t(c|x))$ remains sufficiently low. We experimentally found 0.6 to work consistently well across models and datasets.

As shown in Table 1, an average accuracy gain of 5.0% over k-means is found for CIFAR100 and ImageNet, even with the plain PMI setup. Introducing multiple heads in PMI further improves the obtained results by an average gain of 0.8%. Critically, for our best setup (TEMI) we observe an average gain of 8.1% and 3.7% compared to k-means and the SCAN clustering loss, respectively. Note that even 16 heads were sufficient to get similar performance, specifically less than 1% accuracy deterioration compared to 50 heads.
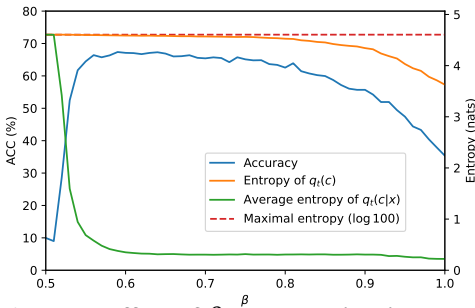
Figure 2: **Effect of β on the validation accuracy and on the entropy of $q_t(c|x)$ and $q_t(c)$ on CIFAR100**. The values are computed using MWMI DINO ViT-B/16. The dashed horizontal line illustrates the maximal possible entropy, i.e. $\log C$. A high entropy of $q_t(c)$ indicates that the clusters are almost uniformly utilized, while a low entropy of $q_t(c|x)$ indicates highly confident predictions (one-hot).
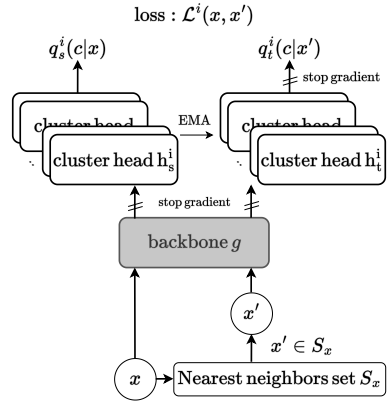

Figure 3: **An overview of the proposed self-distillation clustering framework.** The nearest neighbors are mined in the feature space of g. EMA refers to the exponential moving average over the parameters.

Method	Heads	CIFAR100	ImageNet
k-means	-	56.99	52.26
SCAN	50	62.6±0.94	55.6±0.15
PMI	1	61.6±0.41	57.5±0.22
WMI	1	63.4±1.89	56.5±0.41
PMI	50	63.1±0.56	57.7±0.06
WPMI	50	65.6±1.04	57.0±0.38
TEMI	50	**67.1±1.30**	**58.4±0.22**

Table 1: **Ablation study for the TEMI objective**. All the experiments were conducted with $\beta = 0.6$, and DINO ViT-B/16 as the backbone model. The clustering accuracy is reported in %.

To study the applicability of our method, we then applied our best setup (TEMI) to various publicly available pretrained models, as shown in Fig 1. Therein, we report an average accuracy gain of 6.1% and 12.1% compared to *k*-means on ImageNet and CIFAR100 across 17 different pretrained models. More specifically, TEMI MSN ViT-L/14 and TEMI DINO ViT-B/16 are the best-performing self-supervised methods on ImageNet (61.6% ACC) and CIFAR100 (67.1% ACC). Moreover, CLIP-based backbones have the highest ACC increase over *k*-means when trained with TEMI, precisely 10.7% on ImageNet and 14.1% on CIFAR100.

Concerning the supervised pretrained models in Fig. 1, we demonstrate that ConvNext-L outperforms ViT-L on ImageNet, precisely by 2.7% on ACC with TEMI. However, the supervised ViT-L surpasses ConvNext-L by a large margin of 22.3% in ACC, when benchmarked on CIFAR100 with TEMI. Among the architectures investigated, large ViTs learn the most transferable label-related features, even without supervised fine-tuning. This finding is consistent with [29].

Clustering and overclustering results on ImageNet. Regarding ImageNet, we compare various self-supervised architectures that were trained without any external data, as depicted in Table 2. Using the same architecture (Resnet50) as current state-of-the-art models (SSCN

Method	Arch.	ACC	ARI
SeLa [53]	Resnet50	30.5	16.2
SCAN [34]	Resnet50	39.9	27.5
SSCN [1]	Resnet50	41.1	29.5
Our method			
TEMI DINO	Resnet50	45.2	31.3
TEMI DINO	ViT-B/16	58.4	45.9
TEMI MSN	ViT-L/16	**61.6**	**48.4**

Table 2: **Clustering results for the ImageNet validation set, without using additional data.** Evaluation metrics include clustering accuracy (ACC), and adjusted random index (ARI) in %. All our models are pretrained on ImageNet.

Method	AMI (%)
DeepCluster [4]	28.1
MoCo [18]	28.5
PCL [25]	41.0
ProPos [21]	52.5
TEMI DINO Resnet50	51.8±0.1
TEMI DINO ViT-B/16	**59.9±0.2**
TEMI MSN ViT-L/16	58.8±0.5

Table 3: **Overclustering results on the ImageNet validation set.** The adjusted mutual information (AMI) score for 25K clusters is reported, as in [25]. For all experiments, we set $\beta = 1$.

[1]), TEMI achieves an improvement of 4.1% in ACC. With MSN ViT-L/16 as the backbone, we push the state-of-the-art ACC on ImageNet to 61.6%, resulting in a substantial gain of 20.5% compared to SSCN. The obtained results strongly indicate that first learning the augmentation-invariant features and then focusing on learning the invariances w.r.t. images that belong to the same class is an effective strategy for image clustering. Incentivized by the above observation, we investigate the overclustering performance in Table 3, by adopting the setup from [25]. More concretely, we use 25K clusters and set $\beta = 1$ as in ProPos. We almost match the performance of ProPos [21] with TEMI DINO Resnet50 without tuning the number of clusters or any other hyperparameter, while reaching a considerable gain of 7.4% in AMI with TEMI DINO ViT-B/16.

Small-scale benchmarks. In Table 4, the transfer performance on three small-scale datasets is evaluated. TEMI DINO ViT-B backbone has inarguably the best transfer performance, outperforming the ACC of ProPos by 4.6% and TSP by 2.9% on average. It is worth pointing out that TSP [40] uses the same pretrained model and it is thus a fair comparison. Ultimately, we notice a large accuracy gap between clustering methods and probing in CIFAR20, which suggests that the superclass structure cannot be inferred from the visual input. For instance, clocks, lamps, and telephones are grouped into household electrical devices.

Analysis on noise (false positives) from the NN of g. As shown in Tab. 5, when keeping only the true positive neighbours from 50-NN, we increased the performance from 67.1 → 82.6 using TEMI DINO ViT-B on CIFAR100. We also show that the head weighting term of TEMI in Eq. 10 is not needed, highlighting that TEMI is designed for the noisy pairs obtained from k-NN. As a reference, DINO ViT-Base has 72% true positive pairs in 20-NN, and 66% in the mined 50-NN on CIFAR100.

4.3   Discussion

How expressive can an image classifier be by only training with NN pairs? We examine the training accuracy in Table 5, by training with the true positive pairs from the computed k-NN. The 98.6% training accuracy on CIFAR100 with TEMI DINO ViT-B/16 indicates that it is possible to train a powerful unsupervised image classifier by only relying on pairs. In

Methods	CIFAR10	CIFAR20	STL10
NNM [[1]]	84.3	47.7	80.8
PCL [[25]]	87.4	52.6	41.0
SCAN [[54]]	88.3	50.7	80.9
SPICE [[50]]	92.6	53.8	93.8
ProPos* [[21]]	94.3	61.4	86.7
TSP[†] [[40]]	94.0	55.6	97.9
TEMI[†]	**94.5**	**63.2**	**98.5**
supervised baseline			
Probing[†]	96.8	89.5	99.2

Table 4: **Clustering accuracy on small datasets.** Methods with [†] use DINO ViT-B pretrained on ImageNet, while * indicates methods that include the validation split during training.

Loss	Val. ACC	Train ACC
PMI	**84.1±0.36**	**98.6±0.38**
TEMI	82.6±0.67	96.5±0.88
supervised baseline		
Probing	85.3	99.3

Table 5: **Clustering accuracies on CIFAR100 when training only with the true positive NN pairs using TEMI DINO ViT-B/16.**

fact, we observe that we almost match the supervised linear probing accuracy on CIFAR100 (84.1% vs 85.3%). Still, we identify cases where the human-annotated label is ambiguous and cannot be determined solely by the visual signal (supplementary material).

What is the impact of the instance weighting term? After training, we examined the actual value of the instance weighting term $w(x,x')$. To this end, we computed the mean weights for true positives and false positives sampled from 50-NN within the CIFAR100 validation set, which take the values 0.76 and 0.40, respectively. Furthermore, $w(x,x')$ has a negative impact when only true positive pairs are considered during training (Table 5). This is an expected behaviour, as a fraction of true positive pairs will be down-weighted by $w(x,x')$ due to low feature similarity (i.e. digital and analogue clocks).

How discriminative are the cluster assignments of TEMI? Besides Fig. 2, we quantify the discriminative power of TEMI by computing the mean and median maximum softmax probability (MSP [[19]]). We calculate a mean and median MSP of 88.5% and 98.9% on CIFAR100 and 85.3% and 99.2% on ImageNet. The computed results verify that the introduced framework results in discriminative predictions.

Joint learning of encoder and cluster head. When jointly training the pretrained backbone with the already trained head, we observed a performance increase only when the pretraining dataset was different from the downstream dataset (67.1% → 70.9% ACC on CIFAR100 using the ImageNet-pretrained DINO ViT-B/16). We hypothesize this enables learning features specific to the training distribution. Still, the structure in the latent space of the pretrained model is required for TEMI to determine the k-NN.

5   Conclusion

In this paper, a novel and general self-distillation framework for image clustering was proposed that can achieve competitive results almost out of the box. In addition, a new objective based on pointwise mutual information was presented. After studying the performance of 17 pretrained models, it was shown that TEMI can be used with any type of pretraining with significant improvements over k-means. Finally, new state-of-the-art results were achieved on ImageNet both for clustering and overclustering, leveraging self-supervised ViTs. To conclude, future works are encouraged to explore the connection between image clustering and representation learning in greater depth.

References

[1] Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022.

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 456–473. Springer, 2022. doi: 10.1007/978-3-031-19821-2_26. URL https://doi.org/10.1007/978-3-031-19821-2_26.

[3] Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[5] 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 IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[6] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, pages 5879–5887, 2017.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[8] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[11] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13693–13702, 2021.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] 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. *arXiv preprint arXiv:2010.11929*, 2020.

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021.

[15] Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5414–5423, 2021.

[16] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] 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.

[19] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[20] Jiabo Huang, Qi Dong, Shaogang Gong, and Xiatian Zhu. Unsupervised deep learning by neighbourhood discovery. In *International Conference on Machine Learning*, pages 2849–2858. PMLR, 2019.

[21] Zhizhong Huang, Jie Chen, Junping Zhang, and Hongming Shan. Learning representation for clustering via prototype scattering and positive sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[22] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[23] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[25] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020.

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[27] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[28] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[29] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[30] Chuang Niu, Hongming Shan, and Ge Wang. Spice: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278, 2022.

[31] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[34] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *European conference on computer vision*, pages 268–285. Springer, 2020.

[35] Jianlong Wu, Keyu Long, Fei Wang, Chen Qian, Cheng Li, Zhouchen Lin, and Hongbin Zha. Deep comprehensive correlation mining for image clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8150–8159, 2019.

[36] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*, pages 478–487. PMLR, 2016.

[37] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*, pages 3861–3870. PMLR, 2017.

[38] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hyx-jyBFPr.

[39] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[40] Xingzhi Zhou and Nevin L Zhang. Deep clustering with features from self-supervised pretraining. *arXiv preprint arXiv:2207.13364*, 2022.