# Exploring the Limits of Deep Image Clustering using Pretrained Models

Nikolas Adaloglou*, Felix Michels*, Hamza Kalisch, Markus Kollmann

Heinrich-Heine-Universität Düsseldorf, Germany

*Equal contribution

{adaloglou,felix.michels}@hhu.de

## Overview

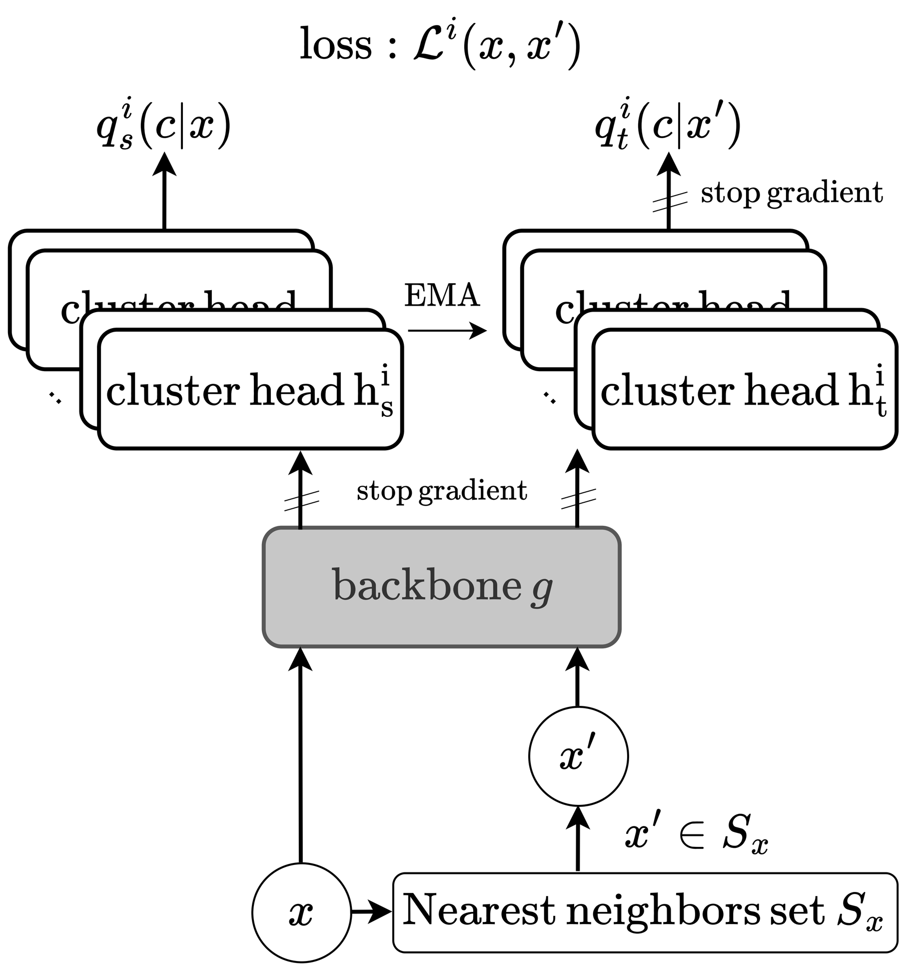
We present a general methodology that learns to classify images without labels by leveraging pretrained feature extractors. We focus on learning the cluster assignments with a novel objective called TEMI, which is based on pointwise mutual information and instance weighting within a multi-head self-distillation clustering framework.

Code: `https://github.com/HHU-MMBS/TEMI-official-BMVC2023`.

### Main Contributions and Findings

- TEMI: A novel and theoretically justified clustering objective with a single **bounded** hyperparameter ($\beta \in (0.5, 1]$).
- Novel clustering framework with **consistent out-of-the-box improvements** across 17 visual backbones and 5 datasets over previous state-of-the-art methods.
- Existing self-supervised ViTs achieve state-of-the-art **clustering accuracy of 61.6% and over-clustering AMI of 59.9% on ImageNet**, without labels or external data.
- ViTs learn the most transferable label-related features when applied to new downstream datasets.

## TEMI: Self-distillation clustering framework




TEMI involves self-distillation training of multiple clustering heads $h$ (3-layer MLPs), based on the fact that nearest neighbors ($x'$ of $x$ from $S_x$) in feature space of $g$ likely share the same semantic label. Cluster predictions are denoted as $q_t^i(c|x)$, $q_s^i(c|x)$ for the teacher $t$ and student $s$ from head $i$. EMA denotes an exponential moving average.

## The pointwise mutual information (PMI) loss

We need to assign an image $x$ to a cluster $c \in \{1, \dots, C\}$. To do this we learn a classifier $q(c|x)$ by maximizing the pointwise mutual information $\mathrm{pmi}(x, x')$ between images of the same class, defined by

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

Under mild conditions, this leads to an optimal solution.

**Thm. 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

$$q^*(c|x) = \arg\max_{q(c|x)} \mathbb{E}_{x,x' \sim p(x,x')}[\mathrm{pmi}(x, x')], \tag{2}$$

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.

## Derivation of the TEMI loss

❶ Approximate the **PMI** using the EMA $\tilde{q}_t(c)$ over $q_t(c|x')$. Introduce hyperparameter $\beta$ to balance class utilization:

$$\mathcal{L}^i(x, x') = -\log \sum_{c=1}^{C} \frac{\left(q_s^i(c|x)q_t^i(c|x')\right)^\beta}{\tilde{q}_t^i(c)}, \tag{3}$$

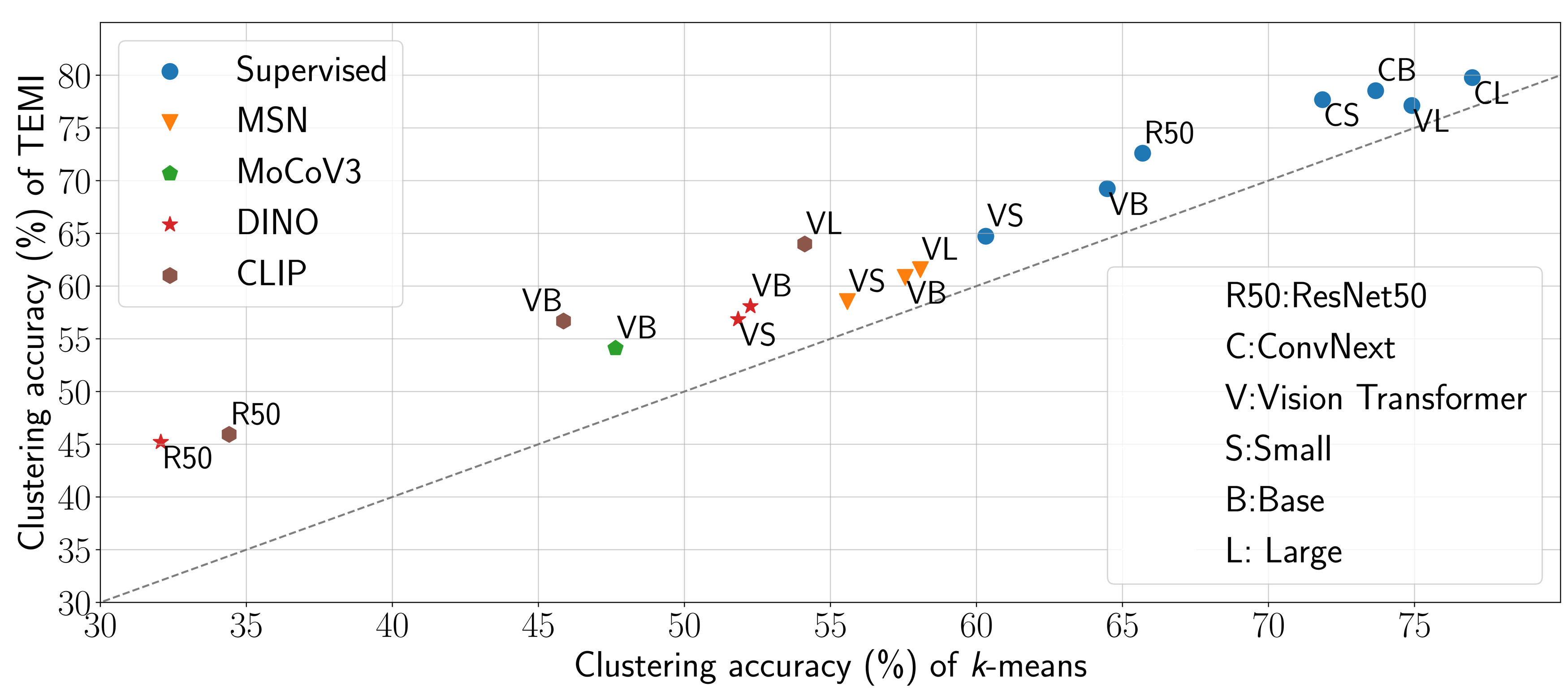
❷ Instance Weighted PMI (**WMPI**) using $q_t^i(c|x)$ to down-weight false positive pairs for each independent head $i$:

$$\mathcal{L}^i_{\mathrm{WPMI}}(x, x') \coloneqq \underbrace{\sum_{c=1}^{C} q_t^i(c|x)q_t^i(c|x')}_{=:w_i(x,x')} \mathcal{L}^i(x, x'). \tag{4}$$

❸ **T**eacher-**E**nsemble p**MI** (**TEMI**): aggregate $w_j(x, x')$ from multiple heads:

$$\mathcal{L}^i_{\mathrm{TEMI}}(x, x') \coloneqq \frac{1}{H} \sum_{j=1}^{H} w_j(x, x')\mathcal{L}^i(x, x'). \tag{5}$$

## Experimental Results




TEMI achieves an **average gain of 6.1% in clustering accuracy compared to $k$-means on ImageNet** across 17 pretrained models. **2.8% improvement on ImageNet when substituting TEMI with SCAN.**

| Method | Arch. | ACC (%) |
|---|---|---|
| SeLa | Resnet50 | 30.5 |
| SCAN | Resnet50 | 39.9 |
| SSCN | Resnet50 | 41.1 |
| *Our method* | | |
| TEMI DINO | Resnet50 | 45.2 |
| TEMI DINO | ViT-B/16 | 58.4 |
| TEMI MSN | ViT-L/16 | **61.6** |

Table 1: **Clustering accuracy in % (ACC) for the ImageNet validation set.**

| Method | Heads | CIFAR100 | ImageNet |
|---|---|---|---|
| $k$-means | - | 57.0 | 52.3 |
| SCAN* | 50 | 62.6 | 55.6 |
| PMI | 1 | 61.6 | 57.5 |
| WPMI | 1 | 63.4 | 56.5 |
| PMI | 50 | 63.1 | 57.7 |
| WMI | 50 | 65.6 | 57.0 |
| TEMI | 50 | **67.1** | **58.4** |

Table 2: **Ablations with DINO ViT-B/16. ACC is reported.**

## Discussion

- **How expressive can a model be just by training with k-NN pairs?** By training with the true positive pairs from the 50-NN, we report 98.6% and 84.1% training and validation accuracy on CIFAR100, which is only 1.2% lower compared to probing, **validating Theorem 1**.
- **Impact of instance weighting.** After training, $w(x, x')$ has a mean value of 0.76 and 0.4 for the true and false positives.
- **How discriminative are the cluster assignments of TEMI?** We calculate a median max softmax probability of 99.2% on ImageNet.