Contrastive Tuning: A Little Help to Make Masked Autoencoders Forget

Johannes Lehner [*1], **Benedikt Alkin** [*1], **Andreas Fürst** [1], **Elisabeth Rumetshofer** [1], **Lukas Miklautz** [2, 3], **Sepp Hochreiter** [1, 4]

[1]ELLIS Unit Linz and LIT AI Lab, Institute for Machine Learning Johannes Kepler University, Linz, Austria
[2]Faculty of Computer Science, University of Vienna, Vienna, Austria
[3]UniVie Doctoral School Computer Science, University of Vienna
[4]Institute of Advanced Research in Artificial Intelligence (IARAI)
lehner@ml.jku.at, alkin@ml.jku.at, lukas.miklautz@univie.ac.at

Abstract

Masked Image Modeling (MIM) methods, like Masked Autoencoders (MAE), efficiently learn a rich representation of the input. However, for adapting to downstream tasks, they require a sufficient amount of labeled data since their rich features code not only objects but also less relevant image background. In contrast, Instance Discrimination (ID) methods focus on objects. In this work, we study how to combine the efficiency and scalability of MIM with the ability of ID to perform downstream classification in the absence of large amounts of labeled data. To this end, we introduce Masked Autoencoder Contrastive Tuning (MAE-CT), a sequential approach that utilizes the implicit clustering of the Nearest Neighbor Contrastive Learning (NNCLR) objective to induce abstraction in the topmost layers of a pre-trained MAE. MAE-CT tunes the rich features such that they form semantic clusters of objects without using any labels. Notably, MAE-CT does not rely on hand-crafted augmentations and frequently achieves its best performances while using only minimal augmentations (crop & flip). Further, MAE-CT is compute efficient as it requires at most 10% overhead compared to MAE pre-training. Applied to large and huge Vision Transformer (ViT) models, MAE–CT excels over previous self-supervised methods trained on ImageNet in linear probing, k-NN and low-shot classification accuracy as well as in unsupervised clustering accuracy. With ViT-H/16 MAE–CT achieves a new state-of-the-art in linear probing of 82.2%. Project page: github.com/ml-jku/MAE-CT.

Introduction

Self-supervised learning (SSL) leverages a pre-training task on unlabeled data to construct rich representations of the input without explicit supervision from costly annotated labels. This pre-trained representation can then be used to solve downstream tasks, like image classification, better than supervised training only. Therefore, SSL is currently one of the most effective machine learning concepts.

Two of the most prominent SSL pre-training tasks in computer vision are *Instance Discrimination* (ID) and *Masked Image Modeling* (MIM). ID uses augmentations to create multiple views of an image. The objective is then to align the views created from the same image. To avoid the trivial solution of mapping all images to a constant representation, methods either use a contrastive loss term (He et al. 2020; Chen et al. 2020), a regularization term (Zbontar et al. 2021) or perform self-distillation (Grill et al. 2020; Caron et al. 2021). When a contrastive loss term is used, the ID task can be viewed as a classification task where each image is its own class. MIM methods, like Masked Autoencoders (MAE) (He et al. 2022) and others (Bao et al. 2022; Xie et al. 2022; Baevski et al. 2022) first mask out areas of the input and then reconstruct the missing parts as pre-train task.

The respective pre-training tasks, classification and reconstruction, of ID and MIM result in distinct advantages and disadvantages. MAE provides a computationally efficient way to exploit sparse pre-training of Vision Transformers (ViT) (Dosovitskiy et al. 2021) by masking large parts of the image (75%) and not processing the masked areas. This computational efficiency, coupled with the data efficiency of a generative reconstruction task (Xie et al. 2023; El-Nouby et al. 2021), enabled beneficial scaling to larger architectures on datasets of limited size. However, to perform well on downstream tasks, MIM methods rely on fine-tuning with a large amount of labeled data as the representation of MIM methods *lack abstraction* after pre-training. In contrast, ID methods learn an object-focused representation (Caron et al. 2021) that typically results in object-specific clusters which is especially useful when few labels are available, as decision boundaries can be drawn much easier between well separated clusters. Furthermore, ID methods notoriously rely on augmentations based on expert knowledge (Tian et al. 2020) to alleviate the problem of *shortcut learning* (Geirhos et al. 2020), which refers to the phenomenon of overfitting to spurious features. MIM suffers less from this issue as all masked parts have to be reconstructed, leaving little room for shortcut solutions. In fact, MAE achieves its best performance when only minimal augmentations (crop & flip) are used.

Given the benefits and downsides of both approaches, an open question remains: What is the best combination of MIM and ID methods to exploit their respective strengths? Namely, use the unlabeled data and compute efficiency of MIM methods to benefit from larger models while also benefiting from the label efficiency of ID methods for good low-shot performance via a meaningful semantic representation.

[*]These authors contributed equally.

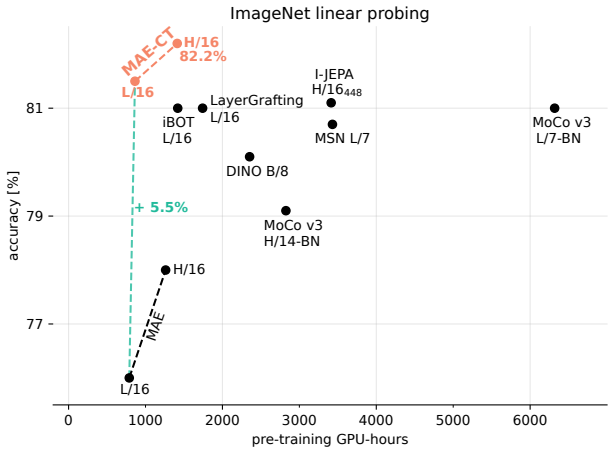

Figure 1: ImageNet linear probing of the best reported models from other self-supervised methods. MAE-CT is able to form a well-separable clustering using little compute.

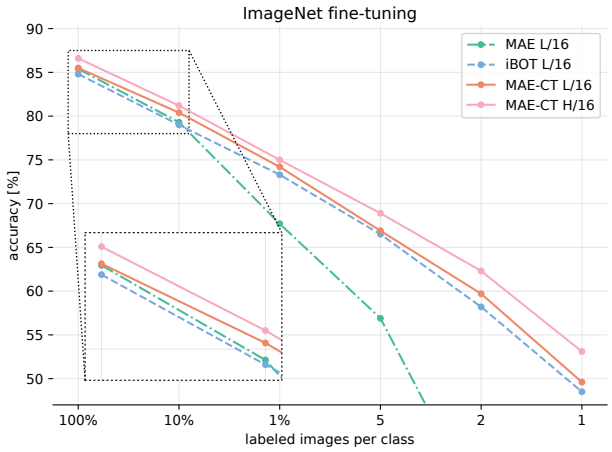

Figure 2: Fine-tuning evaluation on ImageNet. We compare against the best publicly available ID model (iBOT L/16). MAE achieves good performances when given enough labels but struggles otherwise. MAE-CT is able to surpass both MAE and iBOT on all benchmarks. The improvement increases when considering a similar computational budget (MAE-CT H/16) instead of equal model size.

Also, extensive augmentations should be optional as they restrict the field of applications and make the models invariant to potentially useful information. The straightforward combination of MIM and ID in an end-to-end fashion faces the issue that the objectives and hyperparameters are heavily conflicting. For example, MIM benefits from high masking ratios and minimal augmentations while ID benefits from less masking and extensive augmentations.

We propose a sequential self-supervised approach to combine MIM and ID methods named Masked Autoencoder Contrastive Tuning (MAE-CT). Contrastive tuning (CT) aims to imitate fine-tuning in the absence of labeled training samples. MAE-CT utilizes a contrastive objective to guide a pre-trained MAE encoder to forget about the masked pixel-level MAE pre-training objective and to form object-focused semantic clusters. Unlike previous works, MAE-CT uses the contrastive objective not to learn basic features, but to induce abstraction in the top-half layers of the pre-trained ViT model. This novel setting benefits from multiple adaptations.

State-of-the-art ID methods heavily rely on compute-heavy techniques like multi-crop augmentation (Caron et al. 2020), a momentum encoder (He et al. 2020) and long training schedules. We do not rely on these compute-heavy techniques and CT takes only a small fraction of the pre-training duration, consequently, MAE-CT adds only little overhead to MAE pre-training as depicted in Figure 1.

Nevertheless, we find that the disentanglement of feature learning via MAE pre-training and abstraction via contrastive tuning produces well separated clusters. Thus, MAE-CT enables a more label efficient downstream classification than the best ID methods, see Figure 2.

Finally, we adapt the contrastive method Nearest Neighbor Contrastive Learning (NNCLR) (Dwibedi et al. 2021). NNCLR extends SimCLR(Chen et al. 2020) with a queue that holds feature vectors of past samples and a Nearest Neighbor (NN) lookup operation. We observe that this feature space augmentation in combination with our sequential approach renders the use of extensive input augmentations based on expert knowledge optional rather than mandatory.

We provide the following contributions:

1. We introduce MAE-CT, a novel computationally efficient and scaleable approach, to form object-related clusters in the representations of pre-trained MAE encoders.
2. We demonstrate that our compute efficient sequential approach is able to surpass the label efficiency of state-of-the-art ID methods in downstream classification.
3. We find that combined pre-training with minimal augmentations suffers from short-cut learning, providing further evidence for the need of our sequential approach.

Related Work

Critical to our sequential approach are the two components MAE (He et al. 2022) and NNCLR (Dwibedi et al. 2021) on which we build upon. We refer the reader to the respective publications and provide some additional background information in Supplement C. Hereafter, we discuss the most relevant works that make use of both MIM and ID ideas.

The benefit of extending ID with MIM concepts is well established. iBOT (Zhou et al. 2022) adds an auxiliary per-patch reconstruction task. MSN (Assran et al. 2022) uses masking to improve computational efficiency and augmentation strength. Consequently, these methods report state-of-the-art results in feature and low-shot evaluation. We show that the data efficiency and scaling of MAE enables MAE-CT to exceed these results.

The advances in MIM motivated multiple works that extend MIM methods with ID concepts. CMAE (Huang et al.

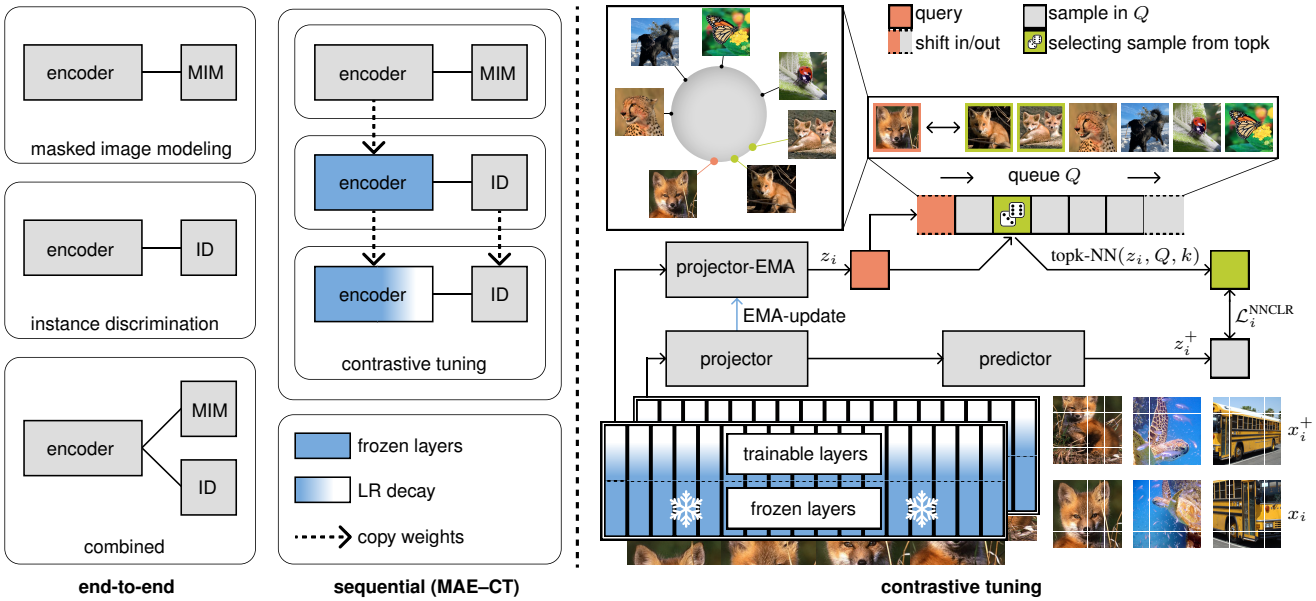

Figure 3: Contrary to end-to-end methods, MAE-CT is a sequential approach, as depicted on the left side. First, an encoder is pre-trained with a MIM objective (MAE). Afterwards, a NNCLR head is initialized on top of said pre-trained encoder by freezing the encoder and training the NNCLR head until its latent representation is well structured. Finally, contrastive tuning is applied for a short duration. In contrastive tuning, depicted on the right side, we freeze the bottom half of the ViT-encoder and apply a layer-wise learning rate decay for the top half. Two views of one image are generated and then encoded by the ViT. Both encodings are fed into a projector, followed by either a predictor or a topk-NN lookup resulting in the embeddings for the NNCLR loss. The queue Q is updated with the new embeddings in a first in – first out manner after each gradient update step.

2022) and BootMAE (Dong et al. 2022) study the addition of an auxiliary ID objective to MAE pre-training. Both approaches improve the fine-tuning performance of ViT-B models, but at the cost of decreased computational efficiency and increased reliance on expert augmentations. Thus, no results on larger models are provided, which would be of interest to measure scalability and data efficiency.

Most related to our approach is the recent work Layer Grafting (Jiang et al. 2023) which also combines MIM and ID into a sequential approach. MAE pre-training is followed by the full training routine of MoCo v3 (Chen, Xie, and He 2021) with an additional regularization loss to keep the lower layer weights close to the MAE weights. In contrast to Layer Grafting, MAE-CT prepares the ID component before changes to the encoder are made, which is then followed up by a short tuning phase. Thus, MAE-CT is much more compute efficient (see Figure 1). For example, the ID part of Layer Grafting requires 15 times the compute of contrastive tuning with a ViT-L/16.

Method

"An autoencoder wants to remember everything a classifier wants to forget." (Epstein and Meir 2019)

Motivation. MIM approaches are able to train large ViT models, that lack the inductive bias of convolutional neural networks, and learn rich representations just from ImageNet (He et al. 2022; Baevski et al. 2022; Xie et al. 2022; Singh et al. 2023). However, MIM models rely on adaptation to the downstream tasks using supervised training where performances heavily degrades as the number of labeled samples decreases. Conversely, ID methods suffer less from this problem as their objective implicitly forms object-specific representations during pre-training(Caron et al. 2021; Walmer et al. 2023). This makes the transition to the downstream task easier as the embedding already represents similar objects in a similar way. The difference in structure can also be seen by evaluating the embedding directly via linear probing or k-NN classification where ID performs significantly better than MIM. Additionally, MAE-CT is motivated by the reported effectiveness of partial fine-tuning (He et al. 2022). Partial fine-tuning improves classification performance considerably by retraining only a few of the topmost layers in a pre-trained MAE with a supervised objective. This implies that features in the lower layers of a pre-trained MAE already generalize well and that only upper layers need to be tuned. Further, fine-tuning induces an object-specific clustering in the representation of the MAE due to label supervision and achieves invariance to certain input features using a set of extensive input augmentations based on expert domain knowledge. Finally, fine-tuning adjusts the model from masked inputs used during MAE pre-training to unmasked inputs which are used in downstream tasks. MAE-CT imitates fine-tuning without labels.

Overview. MAE-CT is a sequential self-supervised approach to induce abstraction in the representation of a

pre-trained MAE. As shown on the left half of Figure 3, MAE-CT requires three steps. First, we perform MAE pre-training. Second, we replace the decoder with a NNCLR head. In this initialization step, the NNCLR head is trained to form fine-grained clusters in the representation that is used for the NN-lookup operation. For this step, the encoder is fully frozen. Third, the Contrastive Tuning (CT) step, where training effects the weights of the upper half of the encoder and the NNCLR head, depicted on the right side of Figure 3 in more detail. During CT the abstract structure of the initialized NNCLR head is transferred back into the encoder to induce a well separated clustering in the encoder output representation.

MAE pre-training MAE pre-training follows the original work (He et al. 2022) to learn a rich but coarsely structured representation in a compute efficient manner by randomly masking out a large fraction of the input patches. We do not apply any masking in the subsequent steps.

NNCLR initialization Like supervised fine tuning, we want to substantially change the encoder representation within a short tuning duration. Accordingly, we find it essential to learn a good target structure in the NNCLR head before changing the encoder.

This is achieved by freezing the encoder and then training only the small fully-connected NNCLR network. Notably, we observe that the NNCLR head is able to map the coarse clusters in the frozen MAE encoder to more fine-grained clusters. We explain this by the observation that, even with the encoder fully frozen and a lightweight NNCLR head, the contrastive objective is able to form a representation with high uniformity in the NNCLR head despite low uniformity in the encoder representation.

As discussed in (Wang and Isola 2020), the contrastive objective can be formulated as the combined minimization of an alignment loss and a uniformity loss. Where alignment is measured as the distance between two views from the same instance and uniformity corresponds to the separability of different instances. Furthermore, the beneficial effect of mapping to a more uniform representation is also reported in (Trosten et al. 2023).

Contrastive tuning Contrastive tuning (CT) uses the initialized NNCLR head to retrain the *partially frozen* MAE encoder. Although we aim to substantially change the encoder representation within a short duration, change has to happen gradually and has to be restricted such that the learned structure in the initialized NNCLR head is not broken. We achieve this by employing layer-wise learning rate decay (Clark et al. 2020) in the encoder. To make the evolution of the entries written into the NNCLR queue more smooth, we use an exponential moving average (EMA) (Laine and Aila 2017; Tarvainen and Valpola 2017) for the lightweight projector network only (Pham et al. 2022).

To reduce memory and compute requirements, we mimic partial fine-tuning and freeze the lower half of the encoder.

Furthermore, CT has to take the increase in trainable parameters into account. We observe that while the target structure in the NNCLR representation quickly improves at first, it starts to degrade before the transformation of the encoder representation can be completed.

To delay this degradation, we increase the difficulty of the alignment task. Instead of using the nearest neighbor of a query vector z_i from the NNCLR queue Q for the NN-lookup, we uniformly sample one of the k nearest neighbors. We refer to this adaptation as topk-NN lookup.

$$\text{topk-NN}(z_i, Q, k) := \underset{\{1,..,k\}}{\mathcal{U}} \left(\underset{q \in Q}{\text{topk}}\ z_i \cdot q \right) \tag{1}$$

Let z^+ refer to the predictor path. The positive counterpart z_i^+ is attracted to a topk-NN of an anchor vector z_i, while all other z_j^+ within the batch are repulsed. Using the temperature τ, we then obtain the updated loss function.

$$\mathcal{L}_i^{\text{NNCLR}} = -\log \frac{\exp\left(\text{topk-NN}(z_i, Q, k) \cdot z_i^+ / \tau\right)}{\sum_{j=1}^{n} \exp\left(\text{topk-NN}(z_i, Q, k) \cdot z_j^+ / \tau\right)} \tag{2}$$

We find that topk-NN lookup improves performance during CT, but not during the initialization step and not in the original NNCLR setting (see ablations in (Dwibedi et al. 2021) Table 7). We argue that this is enabled by the high quality of the initialized NNCLR latent representation on top of the pre-trained MAE features. Consequently, we can increase the strength of the data-driven augmentation effect from the topk-NN lookup by using a higher value for k during CT. Furthermore, as topk-NN lookup effectively extends the average distance between query vector and the surrounding vectors in the queue, it adds the potential to merge isolated subclusters, as denser regions in the surroundings of a query vector are selected with a higher probability.

Experiments and Analysis

Evaluation

We evaluate our approach via image classification on ImageNet (Deng et al. 2009), where we vary the number of used labels from 100% down to a single label per class. To ensure a fair comparison, we exclude results that are based on additional training data or larger sequence lengths (via higher input resolution or smaller patch size). As a lot of large-scale models are not publicly available, we compare MAE-CT to the reported results in Supplement D Table 13.

We evaluate MAE-CT using only minimal image augmentations (MAE-CT$_{min}$) and using the same augmentations as in BYOL (Grill et al. 2020) (MAE-CT$_{aug}$).

We choose the evaluation protocol based on the number of available labels in accordance to previous works. For evaluating the representation using 100% of the labels, we train a linear probe and a k-NN classifier. With 10% and 1% of the labels, we fine-tune the encoder and in the extreme low-shot settings (<1% labels), we report the accuracy of a logistic regression classifier averaged over three splits. The detailed protocols can be found in Supplement B.

Implementation Details

We outline the most important implementation details and provide all further information in Supplement A.

		low-shot evaluations					*feature evaluations*	
Architecture	Method	1 shot	2 shot	5 shot	1%	10%	Linear probing	k-NN
ViT-B/16	MAE (He et al. 2022)	14.0	27.1	43.1	54.2	73.4	68.0	51.1
	MoCo v3 (Chen, Xie, and He 2021)	37.4	47.7	57.3	63.4	74.7	76.7	72.6
	MSN (Assran et al. 2022)	**50.3**	**58.9**	**65.5**	69.5	75.5	77.7	76.3
	iBOT (Zhou et al. 2022)	45.3	55.5	64.3	**71.0**	77.4	**79.5**	**77.1**
	Layer Grafting (Jiang et al. 2023)	40.0	50.2	59.3	65.5	**77.8**	77.7	75.4
	MAE-CT$_{min}$ (ours)	31.1	38.9	47.8	56.6	73.3	73.5	64.1
	MAE-CT$_{aug}$ (ours)	37.5	47.9	57.3	63.3	74.6	76.9	73.4
ViT-L/16	MAE (He et al. 2022)	14.3	34.9	56.9	67.7	79.3	76.0	60.6
	MSN (Assran et al. 2022)	47.5	55.5	62.5	67.0	71.4	77.3	76.2
	iBOT (Zhou et al. 2022)	48.5	58.2	66.5	73.3	79.0	81.0	78.0
	Layer Grafting (Jiang et al. 2023)	47.8	57.6	65.3	69.3	80.1	81.0	77.3
	MAE-CT$_{min}$ (ours)	**51.8**	**60.3**	66.7	72.6	79.7	80.2	78.0
	MAE-CT$_{aug}$ (ours)	49.6	59.7	**66.9**	**74.2**	**80.4**	**81.5**	**79.1**
ViT-H/16	MAE (He et al. 2022)	9.0	16.4	55.2	70.0	80.8	78.0	61.1
	MAE-CT$_{min}$ (ours)	**53.1**	**62.3**	**68.9**	**75.0**	**81.2**	81.5	79.4
	MAE-CT$_{aug}$ (ours)	50.1	60.2	67.7	**75.0**	81.0	**82.2**	**79.8**

Table 1: Low-shot evaluations for different model sizes on ImageNet. "1 shot" corresponds to 1 label per class. "1%" is approximately "13 shot". Feature evaluations are performed with all labels without changing the ViT model.

MAE pre-training. We train for 1600 epochs with a learning rate of $1.5e-4$ and use the "normalize pixels" variant of the MAE loss, which applies a patch-wise normalization to the target pixels before the mean-squared-error loss.

NNCLR initialization. Following (Dwibedi et al. 2021), we use a 3-layer MLP as projector, a 2-layer MLP as predictor and a queue Q of length 65536. To initialize the NNCLR-head, we train for 20 epochs on the output of the fully frozen pre-trained MAE encoder with a learning rate of $1e-4$, a temperature τ of 0.15 and the default top1-NN lookup.

Contrastive tuning. We use a learning rate of $1e-4$ and apply layer-wise learning rate decay (Clark et al. 2020) with decay factor 0.65 to the upper half of the ViT blocks while freezing the lower half. For MAE-CT$_{min}$, we train ViT-B/L for 20 epochs and ViT-H for 30 epochs. For MAE-CT$_{aug}$ we train ViT-B for 80 epochs and ViT-L/H for 40 epochs.

Results

Feature evaluations. The right column of Table 1 shows that MAE-CT improves the linear separability of MAE features considerably on all model sizes. Even larger gains can be observed when a simple distance based k-NN classifier is used. As ID methods learn to prioritize the extraction of object-related information, we find that CT can not lift a MAE pre-trained ViT-B/16 model to the same performance level. But simply increasing the model capacity to ViT-L/16 enables MAE-CT to outperform said ID methods, leveraging the scaleability of MAE pre-training. With ViT-H/16, our sequential approach even exceeds models that operate on higher image resolutions or smaller patches and achieves state-of-the-art in linear probing (see Figure 1).

Low-shot evaluation. The middle column of Table 1 shows classification accuracy when using only a fraction of the labels. Similar to feature evaluations, MAE-CT shows superior scaling as it outperforms state-of-the-art ID methods on larger models. While extensive augmentations are superior for smaller models and more labels, they become less effective as model size grows and the number of labels decreases. For ViT-L/16 models, MAE-CT$_{min}$ surpasses the performance of methods that use extensive augmentations on the 1 shot and 2 shot benchmark. With a ViT-H/16, MAE-CT$_{min}$ is able to surpass the performance of MAE-CT$_{aug}$ on all low-shot benchmarks. Which indicates that the version without additional augmentations benefits even more from the increased model capacity and model depth.

Clustering analysis. We assess the ability of MAE-CT to form object-specific clusters in two ways. First, we use the cluster accuracy (Yang et al. 2010; Xie, Girshick, and Farhadi 2016) to measure how well the ground truth classes of the validation set of ImageNet can be discovered using the unsupervised k-means clustering algorithm. The cluster accuracy ranges between 0 and 100, where 100 indicates a perfect match with the ground truth. Second, we calculate the silhouette score (Rousseeuw 1987) to quantify the spread and compactness of the ground truth classes. The silhouette score ranges from -100 to 100, with 100 being the best value. Silhouette scores smaller than zero indicate that the clusters are not well separated.

Both cluster accuracy and silhouette score are reported in Table 2. Compared to MAE, MAE-CT shows a large improvement in cluster performance. The silhouette score improves from being negative, finding almost no cluster structure, to being positive showing separated clusters. On ViT-L/16, MAE-CT outperforms all other ID methods even when using only minimal augmentations. To the best of our knowledge, the cluster accuracy of MAE-CT (ViT-H/16) is state-of-the-art on ImageNet, when trained on ImageNet only. Further details are provided in Supplement D.

Method	B/16	L/16	H/16
MAE	13.8 (-5.4)	14.3 (-4.1)	11.1 (-7.6)
MoCo v3	43.0 (4.5)	-	-
MSN	**54.2** (10.4)	45.4 (4.8)	-
iBOT	50.0 (6.7)	52.0 (9.0)	-
MAE-CT$_{min}$	35.3 (1.1)	54.9 (11.0)	**58.0** (8.4)
MAE-CT$_{aug}$	46.2 (4.3)	**56.9** (10.1)	54.8 (7.9)

Table 2: k-means cluster accuracy on ImageNet. Parentheses show the silhouette score w.r.t. the ground truth.

Ablations and Analysis

Ablation. We evaluate the impact of essential CT components in Table 3. Masking during contrastive tuning results in a moderate drop in performance (but enables reproduction on a single GPU). The NNCLR head initialization, by training it on top of frozen encoder features before CT, is essential to the performance gains. As MAE-CT$_{min}$ does not use any augmentations besides crop and flip, the data-driven augmentation of the NN lookup is required. Without NN lookup the performance deteriorates considerably due to shortcut learning.

Method	Probing	k-NN
MAE-CT$_{min}$	**80.2**	**77.4**
apply masking (75%) during CT	79.4	75.3
skip NNCLR head initialization	78.5	68.1
NNCLR without NN lookup	70.6	40.5

Table 3: Ablating study of CT$_{min}$ components with ViT-L/16. Applying masking during CT or skipping the NNCLR head initialization on frozen encoder features results in a performance drop, especially in the k-NN accuracy. The NN-lookup is essential when only crop & flip are used as augmentations.

Combined pre-training of MAE and NNCLR. In addition to the sequential MAE-CT approach, we also investigate the combined pre-training of MAE and NNCLR as depicted in the bottom left of Figure 3. To this end, we train a MAE with a ViT-L/16 encoder by additionally attaching a NNCLR head onto the [CLS] token of the encoder. During training we jointly optimize the MAE and NNCLR objectives where we balance the losses $\mathcal{L} = \mathcal{L}^{\text{MAE}} + \lambda\, \mathcal{L}^{\text{NNCLR}}$ with a λ of 0.001. Note, that we keep the augmentations from MAE, which are crop & flip only.

The upper half of Table 4 shows that the combined pre-training slightly improves over MAE but is far worse than MAE-CT. In addition to just combined pre-training, we investigate the application of CT after combined pre-training. We investigate different ways to initialize the NNCLR head before CT in the lower part of Table 4. For "combined pre-training + CT" we initialize a new NNCLR head by training it on top of frozen encoder features just like we do for MAE-CT. For "combined pre-training + CT$_{\text{skip}}$" we reuse the NNCLR head from combined pre-training directly for CT, which effectively *skips* the reinitialization of the NNCLR head. Additionally, instead of initializing an NNCLR head on top of the frozen encoder, one can train an NNCLR head during MAE training by inserting a stop gradient operation before the NNCLR head ("detached pre-training + CT$_{\text{skip}}$"). The results show that directly using the NNCLR head from pre-training for CT is slightly worse in addition to being more constrained as it requires modification of the MAE training process.

Method	Probing	k-NN
MAE	76.0	60.6
combined pre-training	77.8	66.1
combined pre-training + CT$_{\text{skip}}$	79.7	75.3
detached pre-training + CT$_{\text{skip}}$	80.1	77.0
combined pre-training + CT	80.1	76.7
MAE-CT$_{min}$	**80.2**	**77.4**

Table 4: Comparison of combined pre-training of MAE and NNCLR without and with applying CT.

Overall the sequential approach of MAE-CT is more flexible, more compute efficient and achieves superior performances compared to the combined pre-training, even when CT is applied in addition to combined pre-training.

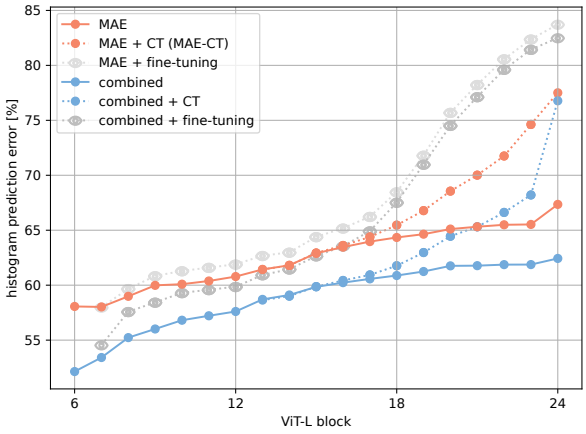

Figure 4: Measuring shortcut learning: Predicting the color histogram of input images from the [CLS] token after different blocks of a ViT-L (higher error is better). Combined pre-training learns shortcut features that can only partially be *forgotten* via CT or supervised fine-tuning. Sequential training avoids shortcut learning.

Shortcut learning of combined pre-training. In experiments with combined pre-training, we observe that even with a small NNCLR loss weight λ during combined pre-training, the NNCLR loss decreases immediately by about 30% compared to training a detached NNCLR head during combined pre-training. As the encoder does not have a good feature representation at the start of training, this indicates that the NNCLR head steers the encoder to extract basic features which drastically simplify the NNCLR objective.

These basic features might be a symptom of shortcut learning (Geirhos et al. 2020). For ID methods, a form of shortcut learning is to learn color statistics of the input image, as two views of the same image likely have a similar color histogram. Following (Addepalli et al. 2022) we utilize a prediction task to estimate to what degree the model preserves information about color statistics within the [CLS] token. We train a linear probe to predict the color histograms of the input image. In the combined pre-training, the NNCLR head pushes the encoder towards learning color statistic features (Figure 4). These shortcut features evolve already in early encoder layers of the combined pre-training. While CT is able to partially correct them, they remain more dominant than in the sequentially trained encoder. Not even supervised fine-tuning of all layers — using augmentations based on expert knowledge — can fully mitigate this effect. We describe the color histogram prediction task in Supplement B.

Cluster formation. To demonstrate the differences in clustering we provide results for ImageNet-Dogs15 (Chang et al. 2017), a subset of ImageNet commonly utilized in the clustering literature. In Figure 5 we show the UMAP (McInnes et al. 2018) embedding of different variations (ViT-L) with their corresponding clustering accuracy. Combined pre-training finds well separated clusters for the classes *Norwegian Elkhound*, *Pug* and *Maltese Dog*. We suspect that these three dog breeds have only small intra-class variations in their characteristics, which makes them easily discernible by low level features. Once contrastive tuning is applied the cluster accuracy improves by a factor of four and also the classes are visually better separated.

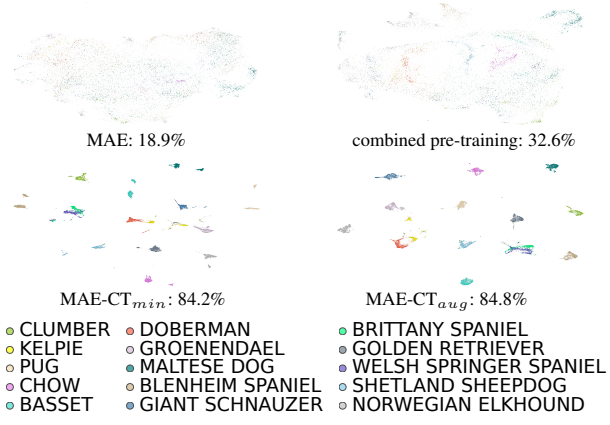

Figure 5: UMAP embeddings of MAE, combined pre-training, MAE-CT$_{min}$ and MAE-CT$_{aug}$ with corresponding k-means cluster accuracies for ImageNet-Dogs15 (ViT-L). MAE-CT clearly improves the separation of the 15 classes.

Cluster retrieval. Figure 6 shows the NNs of two k-means cluster centroids for ImageNet-Dogs15 of MAE and MAE-CT$_{min}$. Inspecting the NNs of the cluster centroid indicates that MAE finds some clusters that correspond to image backgrounds, the first row contains dogs that are located inside and the second row contains dogs that are outside. This is also quantified by the low cluster accuracy w.r.t. the ground truth dog breeds of MAE. MAE reaches a cluster accuracy of 18.7% vs. 94.3% reached by MAE-CT. MAE-CT finds distinct clusters containing mostly images of a single class, shown by the perfect NN retrievals of the classes *Basset* (third row) and *Norwegian Elkhound* (fourth row). Note, that the dogs are correctly grouped despite the different background. We provide the full retrieval, confusion matrices and UMAP embeddings in Supplement E.

Figure 6: Ten NNs for two k-means cluster centers for MAE (upper) and MAE-CT$_{min}$ (lower). Each row corresponds to one cluster found in the [CLS] token of ViT-H/16 for ImageNet-Dogs15. MAE groups the images into dogs located indoors (first row) and outdoors (second row) depending on the background. MAE-CT finds clusters that correspond to the specific dog breeds.

Conclusion

We introduce MAE-CT, a self-supervised approach to combine the strengths of MIM and ID methods. We show that the NNCLR training objective — applied to an already pre-trained MAE model — is capable of creating object-specific clusters in its feature representation which greatly improves representation quality (linear probing, k-NN and cluster accuracy) and low-shot classification performance.

We show that our sequential approach preserves the data efficiency of MAE and incorporates the label efficiency of NNCLR while requiring only 10% more compute than MAE pre-training. This enables the training of large ViT models on ImageNet only, where our larger models exceed the performance of previous state-of-the-art SSL models.

In contrast to state-of-the-art ID methods, MAE-CT does not rely on hand-crafted image augmentation. This is a very promising result, which can be explained by the data-driven augmentation effect of the NN-lookup, which greatly benefits from representations that already capture image semantics in a structured way.

Additional Resources

We provide access to our code, model checkpoints and supplement on our project page: github.com/ml-jku/MAE-CT.

Acknowledgements

We acknowledge EuroHPC Joint Undertaking for awarding us access to Karolina at IT4Innovations, Czech Republic and to MeluXina at LuxProvide, Luxembourg. The ELLIS Unit Linz, the LIT AI Lab, the Institute for Machine Learning, are supported by the Federal State Upper Austria. We thank the projects Medical Cognitive Computing Center (MC3), INCONTROL-RL (FFG-881064), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for GranularFlow (FFG-871302), EPILEPSIA (FFG-892171), AIRI FG 9-N (FWF-36284, FWF-36235), AI4GreenHeatingGrids (FFG- 899943), INTEGRATE (FFG-892418), ELISE (H2020-ICT-2019-3 ID: 951847), Stars4Waters (HORIZON-CL6-2021-CLIMATE-01-01). We thank Frauscher Sensonic, Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, UCB Biopharma SRL, Merck Healthcare KGaA, Verbund AG, GLS (Univ. Waterloo) Software Competence Center Hagenberg GmbH, Borealis, TÜV Austria, TRUMPF and the NVIDIA Corporation.

References

Addepalli, S.; Bhogale, K.; Dey, P.; and Babu, R. V. 2022. Towards Efficient and Effective Self-supervised Learning of Visual Representations. In *European Conference on Computer Vision, ECCV 2022*, volume 13691, 523–538.

Assran, M.; Caron, M.; Misra, I.; Bojanowski, P.; Bordes, F.; Vincent, P.; Joulin, A.; Rabbat, M.; and Ballas, N. 2022. Masked Siamese Networks for Label-Efficient Learning. In *European Conference on Computer Vision, ECCV 2022*, volume 13691, 456–473.

Baevski, A.; Hsu, W.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2022*, volume 162, 1298–1312.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2021*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2021*, 9630–9640.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep Adaptive Image Clustering. In *IEEE International Conference on Computer Vision, ICCV 2017*, 5880–5888.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2022*, volume 119, 1597–1607.

Chen, X.; Xie, S.; and He, K. 2021. An Empirical Study of Training Self-Supervised Vision Transformers. In *IEEE/CVF International Conference on Computer Vision ICCV 2021*, 9620–9629.

Clark, K.; Luong, M.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2009*, 248–255.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; and Yu, N. 2022. Bootstrapped Masked Autoencoders for Vision BERT Pretraining. In *European Conference on Computer Vision, ECCV 2022*, volume 13690, 247–264.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021*.

Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; and Zisserman, A. 2021. With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning of Visual Representations. In *IEEE/CVF International Conference on Computer Vision ICCV 2021*, 9568–9577.

El-Nouby, A.; Izacard, G.; Touvron, H.; Laptev, I.; Jégou, H.; and Grave, E. 2021. Are Large-scale Datasets Necessary for Self-Supervised Pre-training? *CoRR*, abs/2112.10740.

Epstein, B.; and Meir, R. 2019. Generalization Bounds For Unsupervised and Semi-Supervised Learning With Autoencoders. *CoRR*, abs/1902.01449.

Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11): 665–673.

Grill, J.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. Á.; Guo, Z.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 15979–15988.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 9726–9735.

Huang, Z.; Jin, X.; Lu, C.; Hou, Q.; Cheng, M.; Fu, D.; Shen, X.; and Feng, J. 2022. Contrastive Masked Autoencoders are Stronger Vision Learners. *CoRR*, abs/2207.13532.

Jiang, Z.; Chen, Y.; Liu, M.; Chen, D.; Dai, X.; Yuan, L.; Liu, Z.; and Wang, Z. 2023. Layer Grafted Pre-training: Bridging Contrastive Learning And Masked Image Modeling For Label-Efficient Representations. In *The Eleventh International Conference on Learning Representations, ICLR 2023*.

Laine, S.; and Aila, T. 2017. Temporal Ensembling for Semi-Supervised Learning. In *5th International Conference on Learning Representations, ICLR 2017*.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.*, 3(29): 861.

Pham, T.; Zhang, C.; Niu, A.; Zhang, K.; and Yoo, C. D. 2022. On the Pros and Cons of Momentum Encoder in Self-Supervised Visual Representation Learning. *CoRR*, abs/2208.05744.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20: 53–65.

Singh, M.; Duval, Q.; Alwala, K. V.; Fan, H.; Aggarwal, V.; Adcock, A.; Joulin, A.; Dollár, P.; Feichtenhofer, C.; Girshick, R. B.; Girdhar, R.; and Misra, I. 2023. The effectiveness of MAE pre-pretraining for billion-scale pretraining. *CoRR*, abs/2303.13496.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems 30, NeurIPS 2017*, 1195–1204.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What Makes for Good Views for Contrastive Learning? In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*.

Trosten, D. J.; Chakraborty, R.; Løkse, S.; Wickstrøm, K. K.; Jenssen, R.; and Kampffmeyer, M. C. 2023. Hubs and Hyperspheres: Reducing Hubness and Improving Transductive Few-shot Learning with Hyperspherical Embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 7527–7536.

Walmer, M.; Suri, S.; Gupta, K.; and Shrivastava, A. 2023. Teaching Matters: Investigating the Role of Supervision in Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 7486–7496.

Wang, T.; and Isola, P. 2020. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, 9929–9939.

Xie, J.; Girshick, R. B.; and Farhadi, A. 2016. Unsupervised Deep Embedding for Clustering Analysis. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, 478–487.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. SimMIM: a Simple Framework for Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*, 9643–9653.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Wei, Y.; Dai, Q.; and Hu, H. 2023. On Data Scaling in Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, 10365–10374.

Yang, Y.; Xu, D.; Nie, F.; Yan, S.; and Zhuang, Y. 2010. Image Clustering Using Local Discriminant Models and Global Integration. *IEEE Trans. Image Process.*, 19(10): 2761–2773.

Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; and Deny, S. 2021. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, volume 139, 12310–12320.

Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A. L.; and Kong, T. 2022. Image BERT Pre-training with Online Tokenizer. In *The Tenth International Conference on Learning Representations, ICLR 2022*.