What to Hide from Your Students: Attention-Guided Masked Image Modeling

Ioannis Kakogeorgiou[1], Spyros Gidaris[2], Bill Psomas[1], Yannis Avrithis[3,4], Andrei Bursuc[2], Konstantinos Karantzalos[1], and Nikos Komodakis[5,6]

[1]National Technical University of Athens [2]valeo.ai
[3]Institute of Advanced Research in Artificial Intelligence (IARAI) [4]Athena RC
[5]University of Crete [6] IACM-Forth

Abstract. Transformers and masked language modeling are quickly being adopted and explored in computer vision as *vision transformers* and *masked image modeling* (MIM). In this work, we argue that image token masking differs from token masking in text, due to the amount and correlation of tokens in an image. In particular, to generate a challenging pretext task for MIM, we advocate a shift from random masking to informed masking. We develop and exhibit this idea in the context of distillation-based MIM, where a teacher transformer encoder generates an attention map, which we use to guide masking for the student.
We thus introduce a novel masking strategy, called *attention-guided masking* (AttMask), and we demonstrate its effectiveness over random masking for dense distillation-based MIM as well as plain distillation-based self-supervised learning on classification tokens. We confirm that AttMask accelerates the learning process and improves the performance on a variety of downstream tasks. We provide the implementation code at https://github.com/gkakogeorgiou/attmask.

1 Introduction

Self-supervised learning (SSL) has attracted significant attention over the last years. Recently, several studies are shifting towards adapting SSL to transformer architectures. Originating in natural language processing, where self-supervised transformers [14,59] have revolutionized the field, these architectures were introduced to computer vision with the *vision transformer* (ViT) [16] as an alternative to convolutional neural networks [24, 33, 55]. ViT formulates an image as a sequence of tokens obtained directly from raw patches and then follows a pure transformer architecture. Despite the absence of image-specific inductive bias, ViT shows strong image representation learning capacity.

Considering that transformers are data-hungry, many studies advocate pretraining them on unsupervised pretext tasks, determined only by raw data. A

Correspondence: gkakogeorgiou@central.ntua.gr

prominent paradigm is to mask a portion of the input tokens—words in text or patches in images—and train the transformer to predict these missing tokens [2, 14, 23, 66, 71]. This paradigm, called *masked language modeling* (MLM) in the language domain [14], is remarkably successful and extends to the vision domain as *masked image modeling* (MIM) [2, 66, 71].

MIM-based self-supervised methods have already shown impressive results on images. However, an important aspect that has not been well explored so far is how to choose which image tokens to mask. Typically, the selection is random, as has been the norm for text data. In this work, we argue that random token masking for image data is not as effective.

In text, random word masking is likely to hide high-level concepts that describe entire semantic entities such as objects (nouns) and actions (verbs). By contrast, an image has much more tokens than a sentence, which are highly redundant, and random masking is less likely to hide "interesting" parts; or when it does, the remaining parts still easily reveal the identity of the visual concepts. As shown in Figure 1(b-d), unless masking is very aggressive, this is thus less likely to form challenging token reconstruction examples that would allow the transformer to develop strong comprehension skills.

The question we ask is this: *Can we develop a masking strategy that addresses this limitation and makes informed decisions on which tokens to mask?*

To this end, we propose to exploit the intrinsic properties of ViT and in particular its self-attention mechanism. Given an input sequence of image patches, we forward it through the transformer encoder, thereby obtaining an attention map in its output. We then mask the most attended tokens. As shown in Figure 1(f-g), the motivation is that highly-attended tokens form more coherent image regions that correspond to more discriminative cues comparing with random tokens, thus leading to a more challenging MIM task.

This strategy, which we call *attention-guided masking* (AttMask), is an excellent fit to popular distillation-based self-supervised objectives, because it is the teacher encoder that sees the entire image and extracts the attention map, and the student encoder that sees the masked image and solves the reconstruction task. AttMask thus incurs zero additional cost.

We make the following contributions:

1. We introduce a novel masking strategy for self-supervised learning, called AttMask, that exploits the intrinsic properties of ViT by leveraging its self-attention maps to guide token masking (subsection 3.2).
2. We show how to efficiently incorporate this above masking strategy into teacher-student frameworks that use a MIM reconstruction objective and demonstrate significant performance improvements over random masking.
3. Through extensive experimental evaluation, we confirm that AttMask offers several benefits: it accelerates the learning process; it improves performance on a data-limited regime (subsection 4.2) and on a variety of downstream tasks (subsection 4.3); it increases the robustness against background changes, thus revealing that it reduces background dependency.

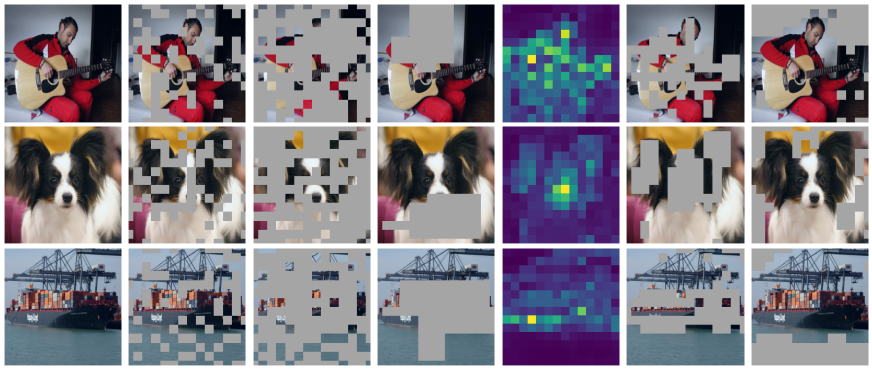


Fig. 1. Different than random masking strategies (b-d), our *attention-guided masking* (AttMask) uses the attention map arising in the encoder (e) to mask the most highly attended by default (f), or the low-attended (g) patches. (b) is used by SimMIM [66], (c) by MAE [23], (d) by BEiT [2] and (g) by MST [36]

2 Related Work

Vision Transformers. Transformers are based on self-attention [59] and require pretraining on large unlabelled corpora [14]. Their adaptation to vision tasks is not straightforward. Representing pixels by tokens is impractical due to the quadratic complexity of self-attention, giving rise to approximations [10, 25, 47, 61, 63]. The idea of representing image patches by tokens is proposed in [12], where patches are of size 2×2, and is further studied in ViT [16], where patches are 16×16. Despite the absence of image-specific inductive bias, ViT is competitive to convolutional neural networks for ImageNet [13] and other smaller benchmark datasets [32, 42]. Since it is pretrained on a large and private dataset [54], authors of DeiT [58] question its efficiency and propose an improved data-efficient version, which however is based on a strong teacher instead [50].

Self-supervised Learning. Early self-supervised learning methods follow the paradigm of training on an annotation-free *pretext task*, determined only by raw data [1, 15, 21, 30, 34, 40, 44, 68]. This task can be *e.g.* the prediction of patch orderings [44] or rotation angles [21]. Starting from instance discrimination [64] and contrastive predictive coding [46], *contrastive learning* has become very popular [3, 8, 17, 28, 39, 53, 62]. These methods pull positives together and push negatives apart, where positives are typically determined by different views of the same example. Alternatively, contrastive learning often relies on clustering [4–6, 19, 35, 67, 72]. The requirement of negatives is eliminated in BYOL [22], OBoW [20], SimSiam [9] and DINO [7], where the challenge is to avoid representation collapse, most notably by a form of *self-distillation* [56].

Using transformers, MIM as a pretext task is proposed in BEiT [2], which maps the images to discrete patch tokens and recovers tokens for masked patches, according to a block-wise random strategy. Other than that, MIM methods use

continuous representations: SimMIM [66] randomly masks large patches and predicts the corresponding pixels by direct regression; MAE [23] randomly masks a large portion of patches and predicts the corresponding pixels using an autoencoder; MST [36] masks low-attended patches and reconstructs the entire input with a decoder; iBOT [71] extends the self-distillation loss of DINO to dense features corresponding to block-wise masked patches. Here, we advocate masking of *highly-attended* patches, in a sense the opposite of MST, and we exhibit this idea in the context of DINO and iBOT.

Regularization and Augmentation. As the complexity of a task increases, networks with more and more parameters are introduced. But with increased representational power comes increased need for more data or risk of overfitting. Several regularization and data augmentation methods have been proposed in this direction [13, 27, 51, 52], combined with standard supervised tasks.

In this context, feature masking is introduced by Dropout [52], which randomly drops hidden neuron activations. To address the strong spatial correlation in convolutional feature maps, SpatialDropout [57] randomly drops entire channels. DropBlock [18] generalizes Dropout—or constrains SpatialDropout—by dropping features in a block, *i.e.*, a square region of a feature map. Attention Dropout [11] makes use of self-attention to mask the most discriminative part of an image. Feature-space masking, guided by attention from another network or branch, has been extensively studied as a mechanism to explore beyond the most discriminative object parts for weakly-supervised object detection [26, 29, 69]. Our work is a natural evolution of these ideas, where attention is an intrinsic mechanism of transformers; and the task becomes that of densely reconstructing the masked features. This is a pretext task, without need for supervision.

3 Method

A simplified overview of the method is shown in Figure 2. We first discuss in subsection 3.1 preliminaries and background on vision transformers and self-supervision with distillation-based masked image modeling. In subsection 3.2, we then detail our attention-guided token masking strategy, called AttMask, and how we incorporate it into masked image modeling.

3.1 Preliminaries and Background

Vision Transformer [16]. We are given an input image $X \in \mathbb{R}^{h \times w \times c}$, where $h \times w$ is the spatial resolution and c is the number of channels. The first step is to tokenize it, *i.e.*, convert it to a sequence of token embeddings. The image is divided into $n = hw/p^2$ non-overlapping patches $P_i \in \mathbb{R}^{p \times p \times c}$ for $i = 1, \ldots, n$, where $p \times p$ is the patch resolution. Each patch is flattened into a vector in $\mathbb{R}^{p^2 c}$ and projected to an embedding vector $\mathbf{z}_i \in \mathbb{R}^d$ using a linear layer, where d is the embedding dimension. A learnable embedding $\mathbf{z}^{[\text{CLS}]} \in \mathbb{R}^d$ of a "classification" token [CLS] is then prepended to form the *tokenized image*

$$Z = (\mathbf{z}^{[\text{CLS}]}; \mathbf{z}_1; \ldots; \mathbf{z}_n) \in \mathbb{R}^{(n+1) \times d}, \tag{1}$$

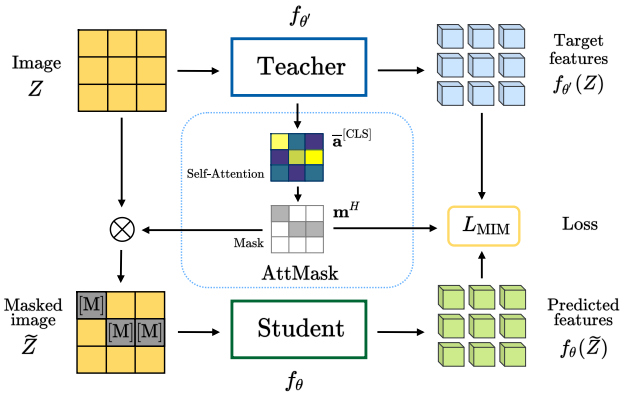


Fig. 2. Simplified overview of AttMask as incorporated in the masked image modelling (MIM) objective of iBOT [71]. A tokenized image Z (1) is given as input to a teacher encoder $f_{\theta'}$, generating target features $f_{\theta'}(Z)$ and an attention map $\bar{\mathbf{a}}^{[\text{CLS}]}$ (7). We then generate a mask $\mathbf{m}^H$ (9) on the most attended tokens and accordingly a masked version $\widetilde{Z}$ (10) of the image, which is given as input to a student encoder f_θ to generate the predicted features $f_\theta(\widetilde{Z})$. Using $\mathbf{m}^H$, loss L_{MIM} (3) is a dense distillation loss between predicted and target features of the masked tokens. Additionally, a global loss L_{G} (4) between [CLS] tokens is applied (not shown here)

where ";" denotes row-wise stacking. The role of this special token is to represent the image at the output. A sequence of position embeddings is added to Z to retain positional information. The resulting sequence is the input to the *transformer encoder*. Each layer of the encoder consists of a *multi-head self-attention* (MSA) block followed by a *multi-layer perceptron* (MLP) block. Through all of its layers, the encoder uses a sequence of fixed length $n+1$ of token embeddings of fixed dimension d, represented by a $(n+1) \times d$ matrix. The embedding of the [CLS] token at the output layer serves as the image representation.

An MSA block consists of a number H of heads, each computing a *scaled dot-product self-attention* [59], *i.e.*, the relevance of each image patch to others, encoded as an $(n+1) \times (n+1)$ *attention matrix*. As discussed in subsection 3.2, we average attention matrices over all the heads of the last encoder layer and we use the row corresponding to the [CLS] token to generate token masks.

Distillation-based Masked Image Modeling. *Self-distillation*, using a moving average of the student as teacher [56], is studied for self-supervision in BYOL [22] and extended to vision transformers in DINO [7], which applies the distillation loss globally on the [CLS] token. iBOT [71] turns this task into *masked image modeling* (MIM) by applying the loss densely on masked tokens.

Given an input image X tokenized as $Z = (\mathbf{z}^{[\text{CLS}]}; \mathbf{z}_1; \ldots; \mathbf{z}_n)$, a *mask vector* $\mathbf{m} = (m_1, \ldots, m_n) \in \{0,1\}^n$ is generated, giving rise to a *masked tokenized image* $\widetilde{Z} = (\mathbf{z}^{[\text{CLS}]}; \tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_n)$, with

$$\tilde{\mathbf{z}}_i = (1 - m_i) \cdot \mathbf{z}_i + m_i \cdot \mathbf{z}^{[\text{MASK}]} \tag{2}$$

for $i = 1, \ldots, n$, where $\mathbf{z}^{[\text{MASK}]} \in \mathbb{R}^d$ is a learnable embedding of a "mask" token [MASK]. Following the strategy of BEiT [2], the mask vector is generated with random *block-wise* token sampling, that is, defined in terms of random rectangles in the 2D layout of the n tokens as a $(h/p) \times (w/p)$ matrix.

Following DINO [7], the transformer encoder is followed by a head that includes an MLP and scaled softmax, such that output token embeddings can be interpreted as probabilities. We denote by f_θ the mapping that includes the addition of the position embeddings, the encoder and the head, while θ is the set of learnable parameters. Given a tokenized image Z, masked or not, we denote by $f_\theta(Z) \in \mathbb{R}^{(n+1)\times d}$ the output token sequence and by $f_\theta(Z)_i, f_\theta(Z)^{[\text{CLS}]} \in \mathbb{R}^d$ the embedding of the i-th and [CLS] token respectively. The teacher parameters θ' are obtained from the student parameters θ by *exponential moving average* (EMA) according to $\theta' \leftarrow \alpha\theta' + (1-\alpha)\theta$.

For each input image, two standard resolution augmented *global views* are generated, with tokenized images Z^a, Z^b and mask vectors $\mathbf{m}^a, \mathbf{m}^b$. For each view v in $V = \{a, b\}$ and for each masked token, the MIM objective is to minimize the reconstruction loss between the student f_θ output for the masked input $\widetilde{Z}^v$ and the teacher $f_{\theta'}$ output for the non-masked input Z^v:

$$L_{\text{MIM}} = -\sum_{v \in V} \sum_{i=1}^{n} m_i^v f_{\theta'}(Z^v)_i \log(f_\theta(\widetilde{Z}^v)_i). \tag{3}$$

Following DINO [7], a similar loss is applied globally on the [CLS] tokens between the student output for one masked view $\widetilde{Z}^v$ and the teacher output for the other non-masked view Z^u:

$$L_{\text{G}} = -\sum_{(u,v) \in V^2} \mathbb{1}_{u \neq v} f_{\theta'}(Z^u)^{[\text{CLS}]} \log(f_\theta(\widetilde{Z}^v)^{[\text{CLS}]}). \tag{4}$$

Finally, as detailed in the supplementary, a *multi-crop* strategy applies, giving rise to a loss L_{LC} between local crops and global views. The overall loss of iBOT [71] is a weighted sum of L_{MIM} (3) and L_{G} (4) + L_{LC}. DINO itself uses the sum L_{G} (4) + L_{LC} without masking.

3.2 AttMask: Attention-guided Token Masking

Prior MIM-based self-supervised methods use random or block-wise random token masking. In this section we describe our attention-guided token masking strategy, which hides tokens that correspond to the salient regions of an image and thus define a more challenging MIM objective.

Attention Map Generation. Given an input sequence $Y \in \mathbb{R}^{(n+1)\times d}$, a *multi-head self-attention* (MSA) layer uses three linear layers to map Y to the *query* Q_j, *key* K_j and *value* V_j sequences for $j = 1, \ldots, H$, where H is the number of heads, $Q_j, K_j, V_j \in \mathbb{R}^{(n+1)\times d'}$ and $d' = d/H$. Then, it forms the $(n+1) \times (n+1)$ *attention matrix*, where softmax is row-wise:

$$A_j = \text{softmax}\left(Q_j K_j^\top / \sqrt{d'}\right). \tag{5}$$

To generate token masks from any layer of the transformer encoder, we average the attention matrices over all heads:

$$\overline{A} = \frac{1}{H} \sum_{j=1}^{H} A_j. \tag{6}$$

Now, each row of an attention matrix is a vector in $\mathbb{R}^{n+1}$, that corresponds to one token and, excluding the diagonal elements, determines an *attention vector* in $\mathbb{R}^n$ over all other tokens. We focus on the attention vector of the [CLS] token, which comprises all but the first elements of the first row of $\overline{A}$:

$$\overline{\mathbf{a}}^{[\text{CLS}]} = \left(\overline{a}_{1,2}, \overline{a}_{1,3}, \ldots, \overline{a}_{1,n+1}\right), \tag{7}$$

where $\overline{a}_{i,j}$ is the element i, j of $\overline{A}$. This vector can be reshaped to $(h/p) \times (w/p)$ *attention map*, to be visualized as a 2D image, indicating the regions of the input image that the [CLS] token is attending.

Mask Generation: Highly-attended Tokens. There is a permutation $\sigma_\downarrow : \{1, \ldots, n\} \to \{1, \ldots, n\}$ that brings the elements of $\overline{\mathbf{a}}^{[\text{CLS}]}$ in descending order, such that $\overline{a}^{[\text{CLS}]}_{\sigma_\downarrow(i)} \geq \overline{a}^{[\text{CLS}]}_{\sigma_\downarrow(j)}$ for $i < j$, where $\overline{a}^{[\text{CLS}]}_i$ is the i-th element of $\overline{\mathbf{a}}^{[\text{CLS}]}$. Choosing a number $k = \lfloor rn \rfloor$ that is proportional to the total number n of tokens with *mask ratio* $r \in [0, 1]$, we define

$$M^H \coloneqq \{\sigma_\downarrow(i), \ldots, \sigma_\downarrow(k)\} \tag{8}$$

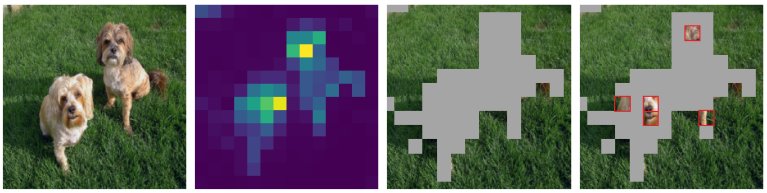
as the set of indices of the top-k most attended tokens. We thus define the *high-attention mask vector* $\mathbf{m}^H$ with elements

$$m^H_i \coloneqq \mathbb{1}_{M^H}(i) = \begin{cases} 1 & \text{if } i \in M^H \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

for $i = 1, \ldots, n$. This masking strategy, which we call AttMask-High, essentially hides the patches that correspond to the most discriminative or salient regions of an image. By AttMask we shall refer to this strategy as default.

Low-attended Tokens. We also examine the opposite approach of AttMask-High that masks the least attended tokens. In particular, we define the set of indices of the bottom-k least attended tokens $M^L = \{\sigma_\uparrow(i), \ldots, \sigma_\uparrow(k)\}$ and the *low-attention mask vector* $\mathbf{m}^L$ with $m^L_i \coloneqq \mathbb{1}_{M^L}(i)$ based on the permutation $\sigma_\uparrow$ that brings the elements of $\overline{\mathbf{a}}^{[\text{CLS}]}$ in ascending order, that is, $\overline{a}^{[\text{CLS}]}_{\sigma_\downarrow(i)} \leq \overline{a}^{[\text{CLS}]}_{\sigma_\downarrow(j)}$ for $i < j$. This strategy, which we call AttMask-Low and is similar to the masking strategy of MST [36], hides patches of the image background. Our experiments show that AttMask-Low does not work well with the considered MIM-based loss.

Highly-attended with Hints. Finally, because AttMask-High may be overly aggressive in hiding the foreground object of an image, especially when the mask ratio r is high, we also examine an alternative strategy that we call AttMask-Hint: While still masking highly attended tokens, we allow a small number of the



(a) Input Image (b) Attention Map (c) AttMask-High (d) AttMask-Hint

Fig. 3. Given image (a), the mean attention map (b) is averaged over heads (6),(7). The AttMask-High strategy (c) masks the most attended patches, while AttMask-Hint (d) reveals few of them to leave hints about the identity of the masked object

most highly attended ones to be revealed, so as to leave hints about the identity of the masked object. In particular, we remove from the initial set M^H a small number $m = \lfloor sn \rfloor$ of tokens with *show ratio* $s < r$. These m tokens are randomly selected from the $\lfloor s_{\max} n \rfloor$ most attended tokens in M^H, where $s_{\max} > s$. An example comparing AttMask-Hint with AttMask-High is illustrated in Figure 3.

Incorporating AttMask into Self-supervised Methods. Because the embedding of the [CLS] token at the output layer of the transformer encoder serves as the image representation, we generate token masks based on the attention vector precisely of the [CLS] token of the output layer. In particular, given a global view tokenized as $Z^v = (\mathbf{z}^{[\text{CLS}]}; \mathbf{z}_1; \ldots; \mathbf{z}_n)$, we obtain the attention vector $\overline{\mathbf{a}}^{[\text{CLS}]}$ (7) and the corresponding high-attention mask vector $\mathbf{m}^H$ (9) at the output layer of the teacher. Then, similarly to (2), we give as input to the student the masked version $\widetilde{Z}^v = (\mathbf{z}^{[\text{CLS}]}; \tilde{\mathbf{z}}_1; \ldots; \tilde{\mathbf{z}}_n)$ with

$$\tilde{\mathbf{z}}_i = (1 - m_i^H) \cdot \mathbf{z}_i + m_i^H \cdot \mathbf{z}^{[\text{MASK}]}. \tag{10}$$

We argue that masking highly attended regions using $\mathbf{m}^H$ helps in learning powerful representations. In section 4, we also experiment with low-attended regions using $\mathbf{m}^L$, supporting further our argument.

AttMask can be incorporated into different methods to either replace the block-wise strategy of BEiT [2] or introduce masking. For iBOT [71], we use $\widetilde{Z}^v$ in L_{MIM} (3) and L_{G} (4). For DINO [7], we introduce masking by using $\widetilde{Z}^v$ for global views in L_{G} (4), but not for local crops in the L_{LC} loss (see supplementary).

4 Experiments

4.1 Setup

Datasets and Evaluation Protocol. We pretrain iBOT and DINO on 20% and 100% of the ImageNet-1k [13] training set. For 20%, we select the first 20% of training samples per class. We evaluate on ImageNet-1k validation set by k-NN or *linear probing*. For linear probing, we train a linear classifier on top of features using the same training protocol as in DINO [7]. With linear probing, we

also validate robustness against background changes on ImageNet-9 (IN-9) [65]. For k-NN [64], we freeze the pretrained model and extract features of training images, then use a k-nearest neighbor classifier with $k = 20$. We also perform the same k-NN experiment, now extracting features only from $\nu \in \{1, 5, 10, 20\}$ examples per class. This task is more challenging and is similar to few-shot classification, only the test classes are the same as in representation learning.

We downstream to other tasks either with or without *finetuning*. We finetune on CIFAR10 [32], CIFAR100 [32] and Oxford Flowers [43] for *image classification* measuring accuracy; on COCO [37] for *object detection* and *instance segmentation* measuring mean average precision (mAP); and on ADE20K [70] for *semantic segmentation* measuring mean Intersection over Union (mIoU). Without finetuning, we extract features as with k-NN and we evaluate using dataset-specific evaluation protocol and metrics. We test on revisited $\mathcal{R}$Oxford and $\mathcal{R}$Paris [49] for *image retrieval* measuring mAP [49]; on Caltech-UCSD Birds (CUB200) [60], Stanford Cars (CARS196) [31], Stanford Online Products (SOP) [45] and In-Shop Clothing Retrieval (In-Shop) [38] for *fine-grained classification* measuring Recall@k [41]; and on DAVIS 2017 [48] for *video object segmentation* measuring mean region similarity $\mathcal{J}_m$ and contour-based accuracy $\mathcal{F}_m$ [48].

In supplementary, we provide more benchmarks, visualizations and ablations.

Implementation Details. As transformer encoder, we use ViT-S/16 [16]. The attention map (7) is generated from the last layer of the teacher encoder by default, *i.e.*, layer 12. We mask the input with probability $p = 0.5$, while the mask ratio r is sampled uniformly as $r \sim U(a, b)$ with $[a, b] = [0.1, 0.5]$ by default. For AttMask-Hint, we set $s_{\max} = 0.1$ and the show ratio s is sampled uniformly from $[s_{\max}a, s_{\max}b] = [0.01, 0.05]$. Following [7, 71], we apply *multi-crop* [6] scheme. The overall loss of iBOT [71] is a weighted sum of L_{MIM} (3), with weight λ, and L_{G} (4) + L_{LC} (DINO [7]), with weight 1, where L_{LC} is the multi-crop loss. By default, $\lambda = 1$. Hyperparameters are ablated in subsection 4.4. Training details are given in the supplementary.

4.2 Experimental Analysis

We provide an analysis on 20% of ImageNet-1k training samples, incorporating AttMask into distillation-based MIM [71] or self-distillation only [7]. We also provide results on robustness against background changes.

Masking Strategies: Distillation-based MIM. We explore a number of masking strategies using distillation-based MIM, by incorporating AttMask into iBOT [71]. We compare AttMask with random block-wise masking [2], which is the default in iBOT, random patch masking with the same ratio, as well as with a more aggressive ratio, following MAE [23]. AttMask masks the most attended tokens (AttMask-High) by default, but we also consider the least attended (AttMask-Low) and the most attended with hints (AttMask-Hint).

We evaluate performance using k-NN and linear probing evaluation protocol on the validation set, along with a fine-tuning evaluation on CIFAR10 and CIFAR100. As shown in Table 1, the AttMask-High outperforms all other masking

Table 1. *Different masking strategies* for iBOT [71] pre-training on 20% of ImageNet. Top-1 accuracy for k-NN, linear probing on ImageNet validation set; fine-tuning on CIFAR10/100. †: default iBOT masking strategy from BEiT [2]. ‡: aggressive random masking strategy from MAE [23]

iBOT Masking	Ratio (%)	ImageNet-1k		CIFAR10	CIFAR100
		k-NN	Linear	Fine-tuning	
Random Block-Wise†	10-50	46.7	56.4	98.0	86.0
Random‡	75	47.3	55.5	97.7	85.5
Random	10-50	47.8	56.7	98.0	86.1
AttMask-Low (ours)	10-50	44.0	53.4	97.6	84.6
AttMask-Hint (ours)	10-50	49.5	57.5	98.1	**86.6**
AttMask-High (ours)	10-50	**49.7**	**57.9**	**98.2**	**86.6**

Table 2. Top-1 k-NN accuracy on ImageNet-1k validation for iBOT pre-training on different percentage (%) of ImageNet-1k. †: default iBOT masking strategy from BEiT [2]

% ImageNet-1k	5	10	20	100
Random Block-Wise†	15.7	31.9	46.7	71.5
AttMask-High (ours)	**17.5**	**33.8**	**49.7**	**72.5**

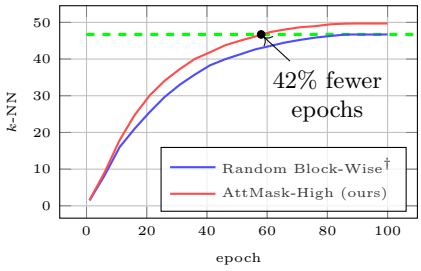


Fig. 4. Top-1 k-NN accuracy on ImageNet-1k validation for iBOT training *vs.* training epoch on 20% ImageNet training set. †: default iBOT masking strategy from BEiT [2]

strategies on all the evaluation metrics. In particular, AttMask-High achieves an improvement of +3.0% on k-NN and +1.5% on linear probing compared with the default iBOT strategy (random block-wise).

Interestingly, random patch masking outperforms the default iBOT strategy, while the more aggressive MAE-like strategy is inferior and AttMask-Low performs the lowest. Intuitively, this means that masking and reconstruction of non-salient regions does not provide a strong supervisory signal under a MIM objective. By contrast, our AttMask creates the more aggressive task of reconstructing the most salient regions and guides the model to explore the other regions. In this setup, AttMask-Hint is slightly lower than AttMask-High.

Data and Training Efficiency. Self-supervised methods on vision transformers typically require millions of images, which is very demanding in computational resources. We advocate that being effective on less data and fast training are good properties for a self-supervised method. In this direction, we assess

Table 3. Top-1 k-NN accuracy on ImageNet-1k validation for DINO [7] pre-training on 20% of the ImageNet-1k training set using mask ratio of 10-50%. †: default DINO

No Masking†	Random	AttMask-Low	AttMask-Hint	AttMask-High
43.0	43.4	42.7	**43.6**	43.5

Table 4. *Background robustness*: Linear probing of iBOT model on IN-9 [65] and its variations, when pre-trained on 20% ImageNet-1k under different masking strategies. †: default iBOT masking strategy from BEiT [2]. ‡: aggressive random masking

iBOT Masking	Ratio (%)	OF	MS	MR	MN	NF	OBB	OBT	IN-9
Random Block-wise†	10-50	72.4	74.3	59.4	56.8	36.3	14.4	15.0	89.1
Random‡	75	73.1	73.8	58.8	55.9	35.6	13.7	14.5	87.9
Random	10-50	72.8	75.3	60.4	57.5	34.9	10.3	14.4	89.3
AttMask-Low (ours)	10-50	66.0	71.1	55.2	52.2	32.4	12.5	14.0	86.6
AttMask-Hint (ours)	10-50	74.4	75.9	61.7	58.3	39.6	**16.7**	**15.7**	89.6
AttMask-High (ours)	10-50	**75.2**	**76.2**	**62.3**	**59.4**	**40.6**	15.2	15.3	**89.8**

efficiency on less data and training time, still with iBOT training. In Table 2 we observe that our AttMask-High consistently outperforms the default random block-wise masking strategy of iBOT at lower percentage of ImageNet-1k training set. In addition, in Figure 4, AttMask-High achieves the same performance as random block-wise with 42% fewer training epochs.

Masking Strategies: Self-distillation Only. Here, we compare masking strategies using distillation only, without MIM reconstruction loss, by incorporating AttMask into DINO [7]. That is, we apply only the cross-view cross-entropy loss on the [CLS] token (4). In Table 3, AttMask-High improves k-NN by +0.5 compared with the default DINO (no masking), while AttMask-Low is inferior. This reveals that AttMask is effective even without a MIM loss. Moreover, AttMask-Hint is slightly better than AttMask-High in this setting.

Robustness Against Background Changes. Deep learning models tend to depend on image background. However, to generalize well, they should be robust against background changes and rather focus on foreground. To analyze this property, we use ImageNet-9 (IN-9) dataset [65], which includes nine coarse-grained classes with seven background/foreground variations. In four datasets, the background is altered: Only-FG (OF), Mixed-Same (MS), Mixed-Rand (MR), and Mixed-Next (MN). In another three, the foreground is masked: No-FG (NF), Only-BG-B (OBB), and Only-BG-T (OBT).

In Table 4, we evaluate the impact of background changes on IN-9 and its variations, training iBOT under different masking strategies. We observe that, except for O.BB. and O.BT, AttMask-High is the most robust. On OBB and

Table 5. Top-1 accuracy on ImageNet validation set. (a) k-NN and linear probing using the full ImageNet training set; (b) k-NN using only $\nu \in \{1, 5, 10, 20\}$ examples per class. Pre-training on 100% ImageNet-1k for 100 epochs

Method	(a) Full		(b) Few Examples			
	k-NN	Linear	$\nu = 1$	5	10	20
DINO [7]	70.9	74.6				
MST [36]	72.1	75.0				
iBOT [71]	71.5	74.4	32.9	47.6	52.5	56.4
iBOT+AttMask-High	72.5	75.7	37.1	51.3	55.7	59.1
iBOT+AttMask-Hint	**72.8**	**76.1**	**37.6**	**52.2**	**56.4**	**59.6**

Table 6. Fine-tuning for *image classification* on CIFAR10 [32], CIFAR100 [32] and Oxford Flowers [43]; *Object detection* (AP^b, %) and *instance segmentation* (AP^m, %) on COCO [37]; and *semantic segmentation* on ADE20K [70] (mIoU, %). Models pre-trained on 100% ImageNet-1k training set for 100 epochs

Method	CIFAR10	CIFAR100	Flowers	COCO		ADE20K
	Accuracy			AP^b	AP^m	mIoU
iBOT	**98.8**	89.5	96.8	48.2	41.8	44.9
iBOT+AttMask	**98.8**	**90.1**	**97.7**	**48.8**	**42.0**	**45.3**

OBT where the foreground object is completely missing, AttMask-Hint exploits slightly better the background correlations with the missing object.

4.3 Benchmark

We pre-train iBOT with AttMask-High and AttMask-Hint on 100% of ImageNet-1k and compare it with baseline iBOT and other distillation-based methods.

ImageNet Classification. As shown in Table 5(a), AttMask-High brings an improvement of 1% k-NN and 1.3% linear probing over baseline iBOT [71] and is better than prior methods. AttMask-High is thus effective for larger datasets too. Table 5(b) shows results of the more challenging task where only $\nu \in \{1, 5, 10, 20\}$ training examples per class are used for the k-NN classifier. In this case, AttMask-High is very effective, improving the baseline iBOT masking strategy by 3-4%, demonstrating the quality of the learned representation. In this setup, AttMask-Hint offers a further small improvement over AttMask-High. For simplicity though, we use AttMask-High by default as AttMask.

Downstream Tasks with Fine-tuning. We fine-tune the pre-trained models with iBOT and iBOT with AttMask for *image classification* on CIFAR10 [32], CIFAR100 [32] and Oxford Flowers [43], *object detection* and *instance segmentation* on COCO [37], and *semantic segmentation* on ADE20K [70]. In Table 6, we observe that AttMask brings small improvement on the baseline iBOT masking

Table 7. *Image retrieval* (mAP, %) on (a) $\mathcal{R}$Oxford and (b) $\mathcal{R}$Paris [49] and *video object segmentation* (mean region similarity $\mathcal{J}_m$ and contour-based accuracy $\mathcal{F}_m$, %) on (c) DAVIS 2017 [48], without fine-tuning. Models pre-trained on 100% ImageNet-1k training set for 100 epochs

Method	(a) $\mathcal{R}$Oxford		(b) $\mathcal{R}$Paris		(c) DAVIS 2017		
	Medium	Hard	Medium	Hard	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
iBOT	31.0	11.7	56.2	28.9	60.5	59.5	61.4
iBOT+AttMask	**33.5**	**12.1**	**59.0**	**31.5**	**62.1**	**60.6**	**63.5**

Table 8. Fine-grained classification (R@k: Recall@k, %) [41] without fine-tuning. Models pre-trained on 100% ImageNet-1k training set for 100 epochs

Method	CUB200			CARS196			SOP			In-Shop		
	R@1	2	4	R@1	2	4	R@1	10	100	R@1	10	20
iBOT	51.4	63.8	75.0	35.6	46.0	56.3	57.4	72.2	84.0	39.1	61.9	68.2
iBOT+AttMask	**57.2**	**69.4**	**80.3**	**39.8**	**50.4**	**61.4**	**59.0**	**73.9**	**85.4**	**40.7**	**63.7**	**70.3**

strategy on *image classification* fine-tuning in all cases. Furthermore, we observe that AttMask improves clearly the scores by 0.6% AP^b on object detection and 0.4% mIoU on semantic segmentation.

Downstream Tasks without Fine-tuning. Without finetuning, we use the pretrained models with iBOT and iBOT with AttMask to extract features as with k-NN and we evaluate using dataset-specific evaluation protocol and metrics. As shown in Table 7(a,b), AttMask is very effective on image retrieval, improving by 1-3% mAP the baseline iBOT masking strategy on $\mathcal{R}$Oxford and $\mathcal{R}$Paris [49], on both medium and hard protocols. More impressive the performance on fine-grained classification, improving by 2-6% R@1 on all datasets, as shown in Table 8. Finally, AttMask improves on video object segmentation on DAVIS 2017 [48] on all metrics, as shown in Table 7(c). These experiments are very important because they evaluate the quality of the pretrained features as they are, without fine-tuning and without even an additional layer, on datasets of different distribution than the pretraining set. AttMask improves performance by a larger margin in this type of tasks, compared with ImageNet.

4.4 Ablation Study

We provide an ablation for the main choices and hyperparameters of our masking strategy and loss function, incorporating AttMask into iBOT [71] and pretraining on 20% of ImageNet-1k training samples.

Layer for Attention Map Generation. The attention map (7) is generated from the last layer of the teacher encoder by default, that is, layer 12 of ViT-S. In Table 9(a), we aim to understand the impact of other layer choices on

Table 9. AttMask k-NN top-1 accuracy on ImageNet-1k validation for iBOT pre-training on 20% of ImageNet-1k *vs*. (a) layer from which the attention map (7) is generated; (b) masking probability p (using batch size 180); and (c) mask ratio r

(a) Layer				(b) Masking Prob p					(c) Mask Ratio r (%)			
6	9	11	12	0	0.25	0.50	0.75	1	10-30	10-50	10-70	30
48.1	48.1	**49.8**	49.7	43.4	47.3	**49.4**	**49.4**	44.2	49.5	**49.7**	48.5	49.1

AttMask. We observe that the deeper layers achieve the highest k-NN performance. Although layer 11 works slightly better, we keep the choice of layer 12 for simplicity, since layer 12 embeddings are used anyway in the loss function.

Masking Probability and Mask Ratio. We mask the global views with probability $p = 0.5$ by default. Table 9(b) reports on other choices and confirms that this choice is indeed best. Therefore, it is useful that student network sees both masked and non-masked images.

The mask ratio r is sampled uniformly as $r \sim U(a, b)$ with $[a, b] = [0.1, 0.5]$ by default. Table 9(c) shows the sensitivity of AttMask with respect to the upper bound b, along with a fixed ratio $r = 0.3$. AttMask is relatively stable, with the default interval $[0.1, 0.5]$ working best and the more aggressive choice $[0.1, 0.7]$ worst. This is possibly due to the foreground objects being completely masked and confirms that masking the most attended patches is an effective strategy. The added variation around the fixed ratio $r = 0.3$ is beneficial.

5 Conclusion

By leveraging the self-attention maps of ViT for guiding token masking, our AttMask is able to hide from the student network discriminative image cues and thus lead to more challenging self-supervised objectives. We empirically demonstrate that AttMask offers several benefits over random masking when used in self-supervised pre-training with masked image modeling. Notably, it accelerates the learning process, achieves superior performance on a variety of downstream tasks, and it increases the robustness against background changes, thus revealing that it reduces background dependency. The improvement is most pronounced in more challenging downstream settings, like using the pretrained features without any additional learning or finetuning, or working with limited data. This reveals the superior quality of the learned representation.

Acknowledgments. We thank Shashanka Venkataramanan for his valuable contribution to certain experiments. This work was supported by computational time granted from GRNET in the Greek HPC facility ARIS under projects PR009017, PR011004 and PR012047 and by the HPC resources of GENCI-IDRIS in France under the 2021 grant AD011012884. NTUA thanks NVIDIA for the support with the donation of GPU hardware. This work has been supported by RAMONES and iToBos projects, funded by the EU Horizon 2020 research and innovation programme, under grants 101017808 and 965221, respectively.

References

1. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 609–617 (2017)
2. Bao, H., Dong, L., Piao, S., Wei, F.: BEit: BERT pre-training of image transformers. In: International Conference on Learning Representations (2022)
3. Cai, T.T., Frankle, J., Schwab, D.J., Morcos, A.S.: Are all negatives created equal in contrastive instance discrimination? arXiv preprint arXiv:2010.06682 (2020)
4. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision. pp. 132–149 (2018)
5. Caron, M., Bojanowski, P., Mairal, J., Joulin, A.: Unsupervised pre-training of image features on non-curated data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2959–2968 (2019)
6. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems **33**, 9912–9924 (2020)
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (2020)
9. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
10. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509 (2019)
11. Choe, J., Shim, H.: Attention-based dropout layer for weakly supervised object localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2219–2228 (2019)
12. Cordonnier, J.B., Loukas, A., Jaggi, M.: On the relationship between self-attention and convolutional layers. In: International Conference on Learning Representations (2020)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. Ieee (2009)
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
15. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1422–1430 (2015)
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)

17. Falcon, W., Cho, K.: A framework for contrastive self-supervised learning and designing a new approach. arXiv preprint arXiv:2009.00104 (2020)
18. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. Advances in Neural Information Processing Systems **31** (2018)
19. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P., Cord, M.: Learning representations by predicting bags of visual words. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
20. Gidaris, S., Bursuc, A., Puy, G., Komodakis, N., Cord, M., Pérez, P.: Obow: Online bag-of-visual-words generation for self-supervised learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
21. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018)
22. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020)
23. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16000–16009 (2022)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
25. Ho, J., Kalchbrenner, N., Weissenborn, D., Salimans, T.: Axial attention in multidimensional transformers. arXiv preprint arXiv:1912.12180 (2019)
26. Hou, Q., Jiang, P., Wei, Y., Cheng, M.M.: Self-erasing network for integral object attention. In: Advances in Neural Information Processing Systems (2018)
27. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456. PMLR (2015)
28. Kalantidis, Y., Sariyildiz, M.B., Pion, N., Weinzaepfel, P., Larlus, D.: Hard negative mixing for contrastive learning. Advances in Neural Information Processing Systems **33**, 21798–21809 (2020)
29. Kim, D., Cho, D., Yoo, D., So Kweon, I.: Two-phase learning for weakly supervised object localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2017)
30. Kolesnikov, A., Zhai, X., Beyer, L.: Revisiting self-supervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1920–1929 (2019)
31. Krause, J., Stark, M., Deng, J., Li, F.F.: 3d object representations for fine-grained categorization. ICCVW (2013)
32. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
33. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. p. 1097–1105. NIPS'12, Curran Associates Inc., Red Hook, NY, USA (2012)
34. Lee, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Unsupervised representation learning by sorting sequences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 667–676 (2017)

35. Li, J., Zhou, P., Xiong, C., Hoi, S.: Prototypical contrastive learning of unsupervised representations. In: International Conference on Learning Representations (2021)
36. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., et al.: Mst: Masked self-supervised transformer for visual representation. Advances in Neural Information Processing Systems **34** (2021)
37. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
38. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
39. Misra, I., Maaten, L.v.d.: Self-supervised learning of pretext-invariant representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6707–6717 (2020)
40. Misra, I., Zitnick, C.L., Hebert, M.: Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. pp. 527–544. Springer (2016)
41. Musgrave, K., Belongie, S., Lim, S.N.: A metric learning reality check. In: European Conference on Computer Vision (2020)
42. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
43. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics and Image Processing (Dec 2008)
44. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on Computer Vision. pp. 69–84. Springer (2016)
45. Oh Song, H., Xiang, Y., Jegelka, S., Savarese, S.: Deep metric learning via lifted structured feature embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
46. Van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv e-prints pp. arXiv–1807 (2018)
47. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International Conference on Machine Learning. pp. 4055–4064. PMLR (2018)
48. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
49. Radenović, F., Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5706–5715 (2018)
50. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollar, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (2015)

52. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(56), 1929–1958 (2014)
53. Stojnic, V., Risojevic, V.: Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1182–1191 (2021)
54. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 843–852 (2017)
55. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9 (2015)
56. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results (2017)
57. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 648–656 (2015)
58. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
59. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
60. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Tech. Rep. CNS-TR-2011-001, California Institute of Technology (2011)
61. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. pp. 108–126. Springer (2020)
62. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
63. Weissenborn, D., Täckström, O., Uszkoreit, J.: Scaling autoregressive video models. In: International Conference on Learning Representations (2020)
64. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018)
65. Xiao, K., Engstrom, L., Ilyas, A., Madry, A.: Noise or signal: The role of image backgrounds in object recognition. In: International Conference on Learning Representations (2021)
66. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9653–9663 (2022)
67. YM., A., C., R., A., V.: Self-labelling via simultaneous clustering and representation learning. In: International Conference on Learning Representations (2020)
68. Zhang, L., Qi, G.J., Wang, L., Luo, J.: Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2547–2555 (2019)

69. Zhang, X., Wei, Y., Feng, J., Yang, Y., Huang, T.S.: Adversarial complementary learning for weakly supervised object localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (June 2018)
70. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017)
71. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. In: International Conference on Learning Representations (2022)
72. Zhuang, C., Zhai, A.L., Yamins, D.: Local aggregation for unsupervised learning of visual embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6002–6012 (2019)