CL-MAE: Curriculum-Learned Masked Autoencoders

Neelu Madan[1,◇], Nicolae-Cătălin Ristea[2,3,◇], Kamal Nasrollahi[1,4],
Thomas B. Moeslund[1], Radu Tudor Ionescu[3,5,*]
[1]Aalborg University, Denmark, [2]University Politehnica of Bucharest, Romania,
[3]University of Bucharest, Romania, [4]Milestone Systems, Denmark, [5]SecurifAI, Romania

Abstract

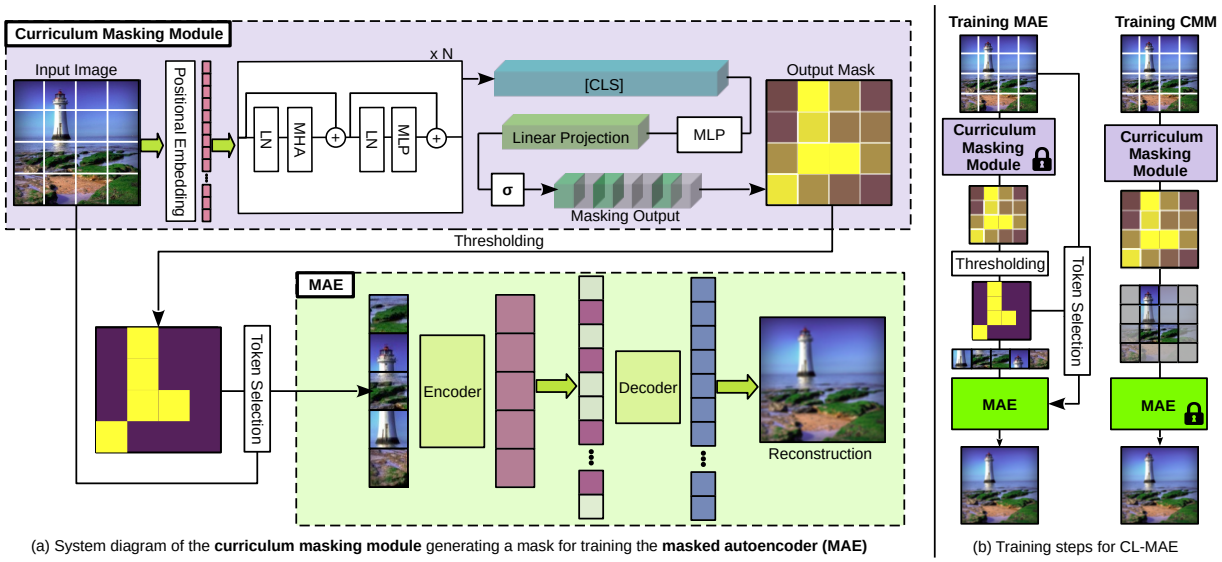
Masked image modeling has been demonstrated as a powerful pretext task for generating robust representations that can be effectively generalized across multiple downstream tasks. Typically, this approach involves randomly masking patches (tokens) in input images, with the masking strategy remaining unchanged during training. In this paper, we propose a curriculum learning approach that updates the masking strategy to continually increase the complexity of the self-supervised reconstruction task. We conjecture that, by gradually increasing the task complexity, the model can learn more sophisticated and transferable representations. To facilitate this, we introduce a novel learnable masking module that possesses the capability to generate masks of different complexities, and integrate the proposed module into masked autoencoders (MAE). Our module is jointly trained with the MAE, while adjusting its behavior during training, transitioning from a partner to the MAE (optimizing the same reconstruction loss) to an adversary (optimizing the opposite loss), while passing through a neutral state. The transition between these behaviors is smooth, being regulated by a factor that is multiplied with the reconstruction loss of the masking module. The resulting training procedure generates an easy-to-hard curriculum. We train our Curriculum-Learned Masked Autoencoder (CL-MAE) on ImageNet and show that it exhibits superior representation learning capabilities compared to MAE. The empirical results on five downstream tasks confirm our conjecture, demonstrating that curriculum learning can be successfully used to self-supervise masked autoencoders. We release our code at https://github.com/ristea/cl-mae.

1. Introduction

Self-supervised representation learning has grown to a prominent research topic, thanks to the possibility of learning representations that can be transferred to multiple visual tasks (referred to as downstream tasks), ranging from image recognition [15, 46, 53] and object detection [5, 58, 61] to semantic segmentation [11, 19, 37, 50]. These generic representations are usually learned by defining a self-supervised task, also known as *pretext task*, where the labels are automatically generated from the available data, requiring no human supervision. Motivated by the achievements of masked language modeling techniques in natural language processing (NLP) [12], the field of computer vision has recently embraced masked image modeling as a self-supervised task [2, 6, 7, 14, 25, 30, 30, 41, 51, 55, 56, 59]. Masked image modeling involves masking a number of patches of an image and tasking the model at learning to reconstruct the masked information based on the remaining visible patches. Masked image models can be divided into two main categories with respect to reconstructing the target either as visual tokens [2, 14, 23, 23, 30, 30, 55, 55, 56, 60] or features [49, 51]. The methods based on predicting masked tokens are the most prevalent ones, mostly because of their simplicity and better generalization capabilities. Even though a lot of attention has been dedicated to refining the pretext task [1, 2, 13, 23, 32, 33, 55, 57], comparatively less attention has been paid to the token selection strategy [6, 30, 31]. The mask selection criteria are often based on semantic object parts [6, 30] or uniform sampling [31]. Unlike existing approaches, we propose to generate adaptive masks with different complexity levels, as part of the learning process, instead of using a single masking strategy. To this end, we propose a novel masking module, which is trained in an end-to-end fashion along with the MAE backbone [23]. We also propose a novel curriculum learning setup, where the complexity of the pretext task is increased from easy-to-hard based on the generated masks, helping MAE to achieve better convergence and learn a more robust representation.

We introduce curriculum learning as the core element of our proposed method, while using MAE [23] as the underlying backbone for representational learning. Curriculum learning [3] operates on the premise that models learn to solve tasks in the increasing order of their complexity, which helps to learn robust representations and enhance generalization capabilities. Different from existing curriculum methods [43], we propose to create a curriculum by generating masks of increasing difficulty during training. To achieve this, we propose a novel masking module that is trained together with the MAE backbone, as shown in Figure 1. In order to generate the progressive masks, from easy to hard, we introduce a curriculum loss function to

*corresp. author: raducu.ionescu@gmail.com; ◇equal contribution.



(a) System diagram of the **curriculum masking module** generating a mask for training the **masked autoencoder (MAE)**

(b) Training steps for CL-MAE

Figure 1. Our Curriculum-Learned Masked Autoencoder (CL-MAE) comprises a learnable masking module that decides what tokens need to be masked at each training iteration. The architecture of our module uses N vision transformer (ViT) [15] blocks based on multi-head attention (MHA), layer normalization (LN) and multi-layer perceptrons (MLPs). The final [CLS] token is passed through an MLP, a linear projection and a sigmoid activation (σ), producing token masking probabilities. The masking module uses an easy-to-hard curriculum learning schedule that transitions smoothly from optimizing the same reconstruction objective as the MAE to an adversarial (opposed) objective. Hence, our masking module generates more or less complex masks, depending on its current objective. Our curriculum masking module (CMM) and the MAE [23] are trained in alternating steps, similar to how generative adversarial networks [20] are trained. During inference, the masking module is removed. Best viewed in color.

train our new masking module, which shares the same objective as the pretext task. The complexity of the generated mask is governed by a factor that decides the weight of our curriculum loss function. First, the weight is set to a positive value in order to generate easy masks and facilitate learning the pretext task. The factor is decreased at every epoch, and even flips from positive values to negative values. When the curriculum loss weight reaches negative values, the masking module learns to increase the complexity of the pretext task by generating hard masks, acting as an adversary to the MAE backbone. Hence, our masking module starts with the same objective as the MAE, but gradually transforms into an adversary during training. This generates an easy-to-hard curriculum for the MAE. The architecture of our learnable curriculum masking module consists of a number of vision transformer (ViT) blocks [15], as illustrated in Figure 1 (a). The masking probabilities are derived from the [CLS] token, after applying a multi-layer perceptron (MLP) and a sigmoid (σ) activation. A thresholding operation transforms the masking probabilities into binary values, which are subsequently used to select the tokens for the MAE. The thresholding operation is required to prevent having a trivial reconstruction task (multiplying the tokens with masking probabilities does not really hide the information, and the MAE can easily learn to rescale the pixel values to their original magnitude). Unfortunately, the thresholding operation also prevents gradient propagation. To overcome this limitation, in each training iteration, we alternate between training the MAE and the masking module, as shown in Figure 1 (b).

We conduct nearest neighbor, linear probing and few-shot linear probing experiments on five downstream image classification tasks, comparing the representation learning capabilities of MAE [23] and CL-MAE, upon self-supervising both models on ImageNet [39]. The empirical results confirm the superior performance of our framework across the entire set of tasks and data sets. Moreover, we present ablation results to illustrate the utility of the various losses and components used by our novel masking module.

Our main contributions are summarized below:

- We introduce curriculum learning into the MAE framework [23] to learn robust representations.
- We propose a novel learnable masking module that is capable of generating adaptive masks, according to the desired complexity level.
- We present comprehensive results on five downstream tasks, showing that our curriculum-learned MAE outperforms MAE by significant margins.

2. Related Work

Self-supervised representation learning. The most popular approaches among state-of-the-art self-supervised methods are based on contrastive learning [8, 21, 24, 35, 47] and

masked image modeling [2, 6, 7, 14, 25, 30, 30, 51, 55, 56, 59, 60]. Contrastive learning is based on pulling positive example pairs closer, while pushing negative pairs farther apart to learn robust representations. Since the number of negative pairs is usually very large, many approaches employ hard negative mining to parse negative image pairs. SimCLR [8] is based on end-to-end training and involves a simple one-to-one comparison with each negative instance. MoCo [24] applies a different parsing technique, employing a momentum encoder to create a dynamic dictionary of negative samples. BYOL [21] depends only on the positive pairs and eliminates the need for negative pairs. All these methods treat an image and its augmented versions as positive pairs, thus relying on heavy augmentation techniques. Masked image modeling represents a relatively simpler approach, where the masked regions of an image are reconstructed based on the visible image content. He *et al*. [23] showed that randomly masking a large number of image patches, *i.e*. 75%, results in a challenging pretext task, which generates a robust representation, eliminating the need for data augmentation.

Masked image modeling. A sizeable amount of research nowadays in self-supervised representation learning is based on masked image modeling [2, 6, 7, 14, 25, 30, 30, 51, 55, 56, 59, 60], which is essentially inspired from masked language modeling, *e.g*. BERT [12]. The mainstream methods based on masked image modeling can be categorized into approaches aiming to reconstruct visual tokens [2, 14, 23, 30, 55, 56, 60] or features [49, 51].

Preliminary studies focused on predicting visual tokens typically rely on an external tokenizer, which creates a visual codebook to reconstruct the target information. BeiT [2] and PeCo [14] are based on generating an offline visual codebook using variational autoencoders. Following DaLL-E [38], iBOT [60] later proposed an online tokenizer based on teacher networks generated via self-distillation. To mitigate the requirement of generating a visual codebook, Wei *et al*. [51] proposed to reconstruct Histogram-of-Oriented-Gradient (HOG) features for the masked region. These approaches are now replaced with more straightforward methods [6, 25, 30, 31, 41, 55, 56] that try to directly reconstruct pixel values. He *et al*. [23] proposed an aggressive masking procedure, randomly hiding 75% of the image patches, which seems to result in a very efficient and effective pretext task to learn robust and generic representations. Xie *et al*. [55] increased the complexity of the pretext task by increasing the patch size and reducing the decoder network to a single layer, as they claim that a harder pretext task leads to a better representation.

A subcategory of methods based on using pixel-wise reconstruction [6, 30, 31] has focused on different masking strategies to obtain robust representations. Li *et al*. [31] proposed a token selection strategy for the pyramid-based ViT, as the random selection of He *et al*. [23] does not seem to work in this case. Li *et al*. [30] proposed semantically-guided masking, a framework that contains two modules, a self-supervised part generator, and a MAE [23] for representation learning. Inspired by Li *et al*. [30], Chen *et al*. [6] unified the part generator and MAE into a single differentiable framework. Different from these approaches, we propose a flexible masking strategy throughout the training process, where the masking depends on the desired complexity of the task, which varies from easy to hard in our case. We also introduce a novel masking module that can easily generate the masks for each level of complexity. Hence, our approach is based on a unique curriculum masking strategy, which is not encountered in existing methods.

Curriculum learning. Curriculum learning, as introduced by Bengio *et al*. [3], is a strategy aimed at organizing input data or tasks in a meaningful order, from easy to hard, to enhance the overall learning outcome. It consists of two main components: a curriculum criterion [3] and a scheduling function [3]. Approaches in curriculum learning can be categorized into easy-to-hard (standard curriculum) and hard-to-easy (anti-curriculum) paradigms [43]. In the easy-to-hard paradigm, tasks are presented to the model in increasing order of complexity [3, 9, 26, 34, 40], while the hard-to-easy paradigm reverses this order [4, 42]. Constructing a curriculum involves using either approaches based on external complexity measures, such as the degree of occlusion and the complexity of the shape [3, 16], or self-paced learning techniques [22, 27, 29, 52], in which the neural network dynamically assesses the difficulty of training samples based on their loss. The scheduling function determines when and how to update the training process and can be categorized as discrete or continuous. Discrete schedulers [3, 44] sort and divide the data into discrete subsets, according to the curriculum criterion. Conversely, continuous schedulers [22, 36] provide a gradually increasing proportion of difficult training samples to the model. Our method incorporates an easy-to-hard continuous scheduler based on the complexity of the pretext task, where the reconstruction error of the model is used as a measure to construct the curriculum. Our novel curriculum learning strategy is deeply integrated within the proposed masking module. The module generates easy masks by hiding tokens with low reconstruction errors and hard masks by hiding tokens with high reconstruction errors. The complexity of the task gradually increases during training from easy-to-hard, and, at some point, the masking module learns to produce extremely hard masks via adversarial training.

3. Method

We propose a curriculum learning approach together with a learnable masking module to train masked autoencoders [23]. The proposed curriculum learning setup is

aimed at achieving a robust representation that can be generalized over multiple data sets and visual tasks. The easy-to-hard curriculum is created through a novel masking module, which learns what tokens to mask in order to make the reconstruction task more or less complex. In the beginning of the training process, the masking module is trained with the same objective as the MAE to make the pretext task easier, *i.e*. it learns to mask tokens that are easy to predict by the MAE. As the training progresses, we reduce the magnitude of the module's objective, essentially letting the module mask tokens at random. Further into the training process, we reverse the objective of the proposed masking module with respect to the MAE, which creates an adversarial learning environment where the MAE continues to learn to reconstruct masked tokens, while the masking module tries to hide the tokens that are difficult to reconstruct. The whole process described above is controlled via a curriculum loss factor λ_{CL} that linearly decreases from a positive value to a negative value. This generates a smooth transition between behaviors of our learnable masking module, initially acting as a partner to the MAE, and gradually transitioning to a neutral state, and later, becoming an adversary to the MAE. Aside from the curriculum learning loss, we enforce the prediction of discriminative and diverse masks via additional loss components. In the following, we describe the proposed masking module and the loss functions used to train our module together with the MAE, in an end-to-end fashion.

3.1. Learnable Masking Module

The core element of our framework is the masking module, a pivotal component in crafting masks of varying difficulty levels for the reconstruction task. Taking an input image $\boldsymbol{I} \in \mathbb{R}^{h \times w \times c}$, we divide it, as identically done for the standard MAE, into $n = (h \cdot w)/p^2$ non-overlapping patches, each patch having $p \times p$ pixels. Further, all patches are flattened and projected via a linear layer into the input tokens. To the existing tokens, we concatenate the learnable [CLS] token $\boldsymbol{C} \in \mathbb{R}^d$, obtaining the input tokens $\boldsymbol{T} \in \mathbb{R}^{(n+1) \times d}$, where d is the token dimension.

The input tokens are processed by the transformer-based module, which is inspired by the Vision Transformer (ViT) architecture [15], obtaining the output tokens $\boldsymbol{T}_{out} \in \mathbb{R}^{(n+1) \times d}$. Considering that the [CLS] token encapsulates information about the entire input image, we extract the output class token $\boldsymbol{C}_{out} \in \mathbb{R}^d$ and process it with a multi-layer perceptron (MLP) followed by a sigmoid activation function, as follows:

$$\mathbf{Z} = \sigma(\text{MLP}(\mathbf{C}_{out})), \tag{1}$$

where σ denotes the sigmoid activation, and $\boldsymbol{Z} \in [0, 1]^n$ is the soft output mask. Each element z_i of $\boldsymbol{Z}$ represents the probability of keeping the i-th input token visible, *i.e*. inferring the corresponding token through the MAE for the current iteration. When training the masking module, the tensor $\boldsymbol{Z}$ is directly multiplied with the input tokens, allowing gradients to pass through our module. However, training the MAE with soft masking probabilities makes the reconstruction task much easier, *i.e*. the only job of MAE is to rescale the values of the softly masked tokens to counter the effect of multiplying the original tokens with the soft masking probabilities. Hence, before training the MAE, we apply a thresholding operation to transform the vector of soft masking probabilities into a binary masking vector $\boldsymbol{Z}^* \in \{0, 1\}^n$. Thanks to the Gaussian and Kullback-Leibler losses (detailed in Section 3.3), the threshold can be fixed to 0.5 without tuning. During inference, the masking module is removed.

3.2. Joint Training Procedure

The output of the masking module is a soft vector with values between 0 and 1. This vector is transformed into a binary vector via a thresholding operation, which prevents gradient propagation (its gradients are equal to zero). Therefore, when we train the MAE backbone by masking input tokens, gradients with respect to the reconstruction loss are not propagated to our masking module. To alleviate this issue, we employ a two-step end-to-end training iteration, which independently updates the MAE and our masking module, similar to how generative adversarial networks are trained [20]. In the first step, we use the binary output $\boldsymbol{Z}^*$ of the frozen masking module to select visible tokens for the MAE backbone, leading to a conventional training step for the MAE. In the second step, we freeze the MAE backbone and train the masking module via $\boldsymbol{Z}$, replacing the thresholding and token selection operations with a multiplication operation. More precisely, instead of selecting which tokens should be passed to the MAE, we multiply all input tokens with the soft output of our masking block, allowing gradient propagation. In this fashion, we can propagate gradients with respect to the reconstruction loss of the MAE, transforming our module into a learnable component. We illustrate both training steps in Figure 1 (b).

3.3. Proposed Loss Functions

To train our learnable masking module, we propose four loss functions that are jointly minimized. We present the four losses and their roles below.

Curriculum loss. Taking inspiration from MAE [23], we employ the mean-squared error (MSE) metric between the normalized per-patch pixels of the reconstructed target ($\hat{\boldsymbol{I}}$) and the input image ($\boldsymbol{I}$) in our curriculum learning framework. The curriculum loss function is given by:

$$\mathcal{L}_{\text{CL}}(\hat{\boldsymbol{I}}, \boldsymbol{I}) = \lambda_{\text{CL}}^{(t)} \cdot (\hat{\boldsymbol{I}} - \boldsymbol{I})^2, \tag{2}$$

where $\lambda_{\text{CL}}^{(t)} \in [-1, 1]$ is linearly decreased at each training

step $t \in \{0, 1, 2, ..., T\}$, as follows:

$$\begin{aligned} \lambda_{\text{CL}}^{(0)} &= 1, \\ \lambda_{\text{CL}}^{(t+1)} &= \lambda_{\text{CL}}^{(t)} - k, \end{aligned} \tag{3}$$

where $k \in [0, 2/T]$ is a tunable decay value, and T is the total number of training iterations. Note that k determines how soon the masking module switches from a consensual objective to an adversarial one. When k is set to the minimum value, there is no adversarial training. For the maximum decay $k = 2/T$, the adversarial training starts halfway into the training process. Depending on k, the value of $\lambda_{\text{CL}}^{(T)}$ can be between 1 and -1. As long as $\lambda_{\text{CL}}^{(t)} > 0$, the masking module tries to minimize $\mathcal{L}_{\text{CL}}$, contributing to simplifying the pretext task. When $\lambda_{\text{CL}}^{(t)}$ becomes negative, the masking module starts maximizing the difference between $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}$, becoming an adversary to the MAE. By decreasing λ_{CL} over time, our loss function constructs a curriculum that progresses from a simple to a challenging reconstruction task, thereby fostering a more effective representation learning process.

Gaussian loss. We incorporate a Gaussian objective into our masking module to enforce discriminative outputs. This objective forces the module to be decisive in picking which tokens should be or not be masked, pushing the masking probabilities away from 0.5, towards 0 or 1. Thus, the soft output vector $\boldsymbol{Z}$ is expected to contain values close to 0 for patches that need to be masked, and values close to 1 for patches that must be kept visible. To achieve this behavior, we employ a Gaussian loss, as follows:

$$\mathcal{L}_{\text{Gauss}} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\boldsymbol{Z} - \mu)^2}{2\sigma^2}\right), \tag{4}$$

where μ denotes the mean, σ corresponds to the standard deviation, and $\boldsymbol{Z}$ represents the output of the masking module. To push the masking probabilities away from 0.5, we set μ to 0.5. To regulate the resulting gradients, we use $\sigma = 0.12$ in all our experiments. By training the masking module with the proposed Gaussian loss, it acquires the capability to determine whether a specific patch should be masked or left unmasked. This objective is meant to minimize the difference between the soft vector $\boldsymbol{Z}$ and the binary vector $\boldsymbol{Z}^*$. Hence, when we switch the thresholding operation applied to $\boldsymbol{Z}$ on and off to alternate between updating the weights of the MAE backbone and those of our module, a low difference between $\boldsymbol{Z}$ and $\boldsymbol{Z}^*$ makes the training steps more consistent with each other.

Kullback-Leibler loss. In our curriculum learning setup, the masking module can easily learn to shortcut the reconstruction task. For example, when the module aims to minimize the reconstruction error, it tends to avoid masking altogether. Conversely, when it behaves as an adversary and aims to maximize the reconstruction error, it tends to mask all the patches. To eliminate such shortcuts, we introduce a new loss function that aims to ensure a fixed number of tokens is always masked, regardless of the complexity of the reconstruction task. To enforce a predetermined masking ratio, we integrate a loss based on the Kullback-Leibler (KL) divergence into our learning framework. Our methodology involves generating distinct distributions for tokens that are masked and those that are visible, all based on a predefined masking ratio. This process includes creating two separate bins: one bin tallies the count of masked tokens, while the other bin keeps track of visible tokens. By establishing these distributions, our aim is to align with the target masking ratio. In order to gauge the difference between the intended distribution and the actual distribution of the outputs, we compute the KL divergence, which quantifies the dissimilarity between the two distributions. The loss based on the KL divergence is defined as follows:

$$\mathcal{L}_{\text{KL}} = m \cdot \log\left(\frac{\hat{m}}{m}\right) + v \cdot \log\left(\frac{\hat{v}}{v}\right), \tag{5}$$

where $\hat{m}$ represents the number of tokens to be masked estimated by the masking module, m denotes the desired number of tokens to be masked (our target masking distribution), $\hat{v}$ signifies the estimated number of visible tokens, and v the desired number of visible tokens. The loss is computed by evaluating the logarithmic ratios of the predicted outputs to the target values, weighted by the respective scale factors m and v. The fractions $\frac{\hat{m}}{m}$ and $\frac{\hat{v}}{v}$ in Eq. (5) provide insights into the match between the predicted outputs and the target values. When the predicted outputs align with the target values, these fractions evaluate to 1, and their logarithm to 0. Hence, the loss value becomes 0. The scale factors m and v ensure that each fraction is appropriately weighted.

Diversity loss. While the random masking process used by MAE [23] inherently generates diverse masks, our learnable masking module might collapse to generating a single mask for all image samples. We thus need to employ a mechanism that ensures data diversity. To this end, we introduce a diversity loss that encourages the generation of different mask configurations. For a mini-batch of p samples, the diversity loss is computed as follows:

$$\mathcal{L}_{\text{div}} = \frac{1}{\frac{p \cdot (p-1)}{2}} \sum_{i=1}^{p} \sum_{j=i+1}^{p} \exp\left(-\|\boldsymbol{Z}_i - \boldsymbol{Z}_j\|^2\right), \tag{6}$$

where $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ are the soft masking vectors corresponding to input images $\boldsymbol{I}_i$ and $\boldsymbol{I}_j$, respectively. Minimizing the proposed diversity loss is equivalent to maximizing the sum of distances between all pairs of soft masking vectors in a mini-batch. The loss is normalized with respect to the number of distinct image pairs in a mini-batch, obtaining a loss value that is independent of the batch size.

Joint Loss. The overall loss function used to optimize the masking module encompasses all the objectives presented so far, namely the curriculum loss ($\mathcal{L}_{\text{CL}}$), the Gaussian loss ($\mathcal{L}_{\text{Gauss}}$), the Kullback-Leibler loss ($\mathcal{L}_{\text{KL}}$), and the diversity

loss ($\mathcal{L}_{\text{div}}$). Formally, the masking module is optimized via the following joint loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CL}} + \lambda_{\text{Gauss}} \cdot \mathcal{L}_{\text{Gauss}} + \lambda_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} + \lambda_{\text{div}} \cdot \mathcal{L}_{\text{div}}, \quad (7)$$

where the hyperparameters $\lambda_{\text{Gauss}} > 0$, $\lambda_{\text{KL}} > 0$ and $\lambda_{\text{div}} > 0$ dictate the contributions of the corresponding loss terms to the overall loss function. Note that the curriculum loss $\mathcal{L}_{\text{CL}}$ does not need a scaling factor, since it already includes one in its definition provided in Eq. (2).

4. Experiments and Results

4.1. Data Sets

ImageNet. The ImageNet bemchmark [39] contains over one million images from 1,000 categories, representing the most popular data set in computer vision.

Aerial Images. The Aerial Images data set (AID) [54] comprises 10K aerial images of 600×600 pixels from 30 distinct categories, collected from Google Earth. The images are divided into 5K for training and 5K for testing.

Airbus Wind Turbines. The Airbus Wind Turbines [17] data set comprises over 357K satellite images of 128×128 pixels, where the task is to classify images with and without wind turbines. We randomly split the data set into 80% for training and 20% for testing.

Architectural Heritage Elements. The Architectural Heritage Elements (AHE) [28] data set encompasses 10 distinct cultural heritage classes. This data set comprises 10,130 training images and 1,404 test images. The resolution of each image is 128×128 pixels.

Sea Animals. The Sea Animals [48] data set comprises images of 23 different sea creatures. This data set contains a total of 13,711 images of distinct resolutions. There are 12,339 images for training and 1,372 images for testing.

Sport Balls. The Sport Balls [10] data set is composed of 15 classes representing various sport balls. It incorporates 7,328 training images and 1,841 test images.

4.2. Experimental Setup

Backbones. To compare the MAE and CL-MAE self-supervised training frameworks, we consider three ViT [15] backbones of different sizes, namely base (ViT-B), large (ViT-L) and huge (ViT-H). These backbones are already available in the official PyTorch repository[1] of MAE, which we employ in our experiments.

Evaluation protocols. We start our experiments by self-supervising MAE [23] and CL-MAE on 200 randomly chosen classes from ImageNet [39]. We then evaluate the learned representations on five downstream data sets, considering multiple evaluation scenarios: nearest neighbor, linear probing, and few-shot linear probing. In the first scenario, we apply a nearest neighbor model based on the Euclidean distance on top of the learned latent space. In the linear probing scenario, we train a Softmax layer on top of the learned encoders. The last scenario is similar to the second one, the only difference being the number of samples per class, which is restricted to a value in the set $\{1, 2, 4, 8, 16\}$. To better assess the power of the self-supervised representations, we refrain from fine-tuning the backbones on the downstream tasks. As evaluation metrics, we report the accuracy for the top-1 and top-5 predictions, denoted as Acc@1 and Acc@5, respectively. For the linear probing and few-shot linear probing protocols, we report the average accuracy rates over three runs for each model. This is not necessary for the nearest neighbors models, since they output deterministic predictions.

[1] https://github.com/facebookresearch/mae

Method	λ_{Gauss}	Acc@1	Acc@5
MAE (baseline) [23]	-	**39.2**	61.5
CL-MAE (no curriculum)	1	35.0	56.7
	2	37.8	59.6
	5	35.1	56.7
	10	38.7	**61.7**
	20	38.2	61.3

Table 1. ImageNet results while tuning the hyperparameter λ_{Gauss} controlling the importance of our Gaussian loss. The results are obtained by nearest neighbor models applied on the self-supervised latent space of MAE and CL-MAE based on ViT-B. The curriculum loss is turned off. The top scores are in bold.

4.3. Hyperparameter Tuning

For the vanilla MAE models, we use the hyperparameters recommended by He *et al*. [23] for ImageNet. When we integrate our masking module, we do not change the recommended hyperparameters for the MAE backbones. However, there are some additional hyperparameters for our learnable masking module, which we tune on the ViT-B architecture. We reuse the hyperparameters established for our masking module on the ViT-L and ViT-H backbones.

Tuning for Gaussian loss. For the moment, we turn off the curriculum learning, and focus on training the masking module to generate masking probabilities close to 0 or 1. Hence, we first tune the hyperparameter λ_{Gauss}, which controls the importance of the Gaussian loss. In Table 1, we present preliminary results on ImageNet with various values for λ_{Gauss} for the CL-MAE based on ViT-B. The empirical results indicate that the Gaussian loss is an important objective for our module, requiring a weight that is ten times greater than the other losses to produce optimal performance. We thus set $\lambda_{\text{Gauss}} = 10$ in the subsequent experiments.

Tuning for Kullback-Leibler loss. Next, we need to make sure that our module masks the right amount of patches, not more nor less. We keep the curriculum loss switched off, and tune the hyperparameter λ_{KL}, which represents the

Method	λ_{KL}	Acc@1	Acc@5
MAE (baseline) [23]	-	39.2	61.5
CL-MAE (no curriculum)	0.1	36.5	59.3
	0.2	37.8	59.9
	0.5	38.2	60.6
	1	**39.5**	**61.9**
	2	39.3	61.7

Table 2. ImageNet results while tuning the hyperparameter λ_{KL} controlling the importance of our Kullback-Leibler loss. The results are obtained by nearest neighbor models applied on the self-supervised latent space of MAE and CL-MAE based on ViT-B. The curriculum loss is turned off. The top scores are in bold.

Method	$\lambda_{\text{CL}}^{(T)}$	Acc@1	Acc@5
MAE (baseline) [23]	-	39.2	61.5
CL-MAE	0	39.5	61.9
	-0.1	**41.4**	**64.5**
	-0.15	40.7	63.6
	-0.2	38.1	60.4

Table 3. ImageNet results while tuning the hyperparameter $\lambda_{\text{CL}}^{(T)}$ (equivalent to tuning k) of our curriculum loss. The results are obtained by nearest neighbor models applied on the self-supervised latent space of MAE and CL-MAE based on ViT-B. The top scores are in bold.

weight for the Kullback-Leibler loss. In Table 2, we present results on ImageNet with various values for λ_{KL} for the CL-MAE based on ViT-B. Values below 1 for λ_{KL} produce considerable performance drops with respect to the MAE baseline. The reported accuracy rates show that the optimal weight for the Kullback-Leibler loss is $\lambda_{\text{KL}} = 1$. We choose this setting ($\lambda_{\text{KL}} = 1$) for the following experiments.

Tuning for curriculum loss. For the curriculum loss, we tune the decay value k with respect to the last weight factor $\lambda_{\text{CL}}^{(T)}$, which specifies the importance of the adversarial objective in the last training iteration T. Since the relation between k and $\lambda_{\text{CL}}^{(T)}$ is bijective, we express the tuning in terms of the more intuitive hyperparameter, namely $\lambda_{\text{CL}}^{(T)}$. We consider values for $\lambda_{\text{CL}}^{(T)}$ between 0 and -0.2 and report the corresponding results in Table 3. Turning off the adversarial training, *i.e.* setting $\lambda_{\text{CL}}^{(T)} = 0$, leads to results that are marginally better than the vanilla MAE. In contrast, using too much adversarial training ($\lambda_{\text{CL}}^{(T)} = -0.2$) seems to actually harm the model. The results show that $\lambda_{\text{CL}}^{(T)} = -0.1$ is the optimal value for the CL-MAE based on ViT-B applied on ImageNet. Therefore, we preserve this value for the subsequent experiments.

Tuning for diversity loss. Another important aspect is the diversity of the generated masks. The results presented so far used the default setting for the weight of the diversity loss, namely $\lambda_{\text{div}} = 1$. However, our module might benefit from a higher emphasis on the diversity of the generated masks. To this end, we consider higher values for λ_{div} and present the corresponding results in Table 4. The reported results indicate that $\lambda_{\text{div}} = 2$ is the optimal choice, suggesting that the diversity of the generated masks is indeed important. We set $\lambda_{\text{div}} = 2$ in the following experiments.

Method	λ_{div}	Acc@1	Acc@5
MAE (baseline) [23]	-	39.2	61.5
CL-MAE	1	41.4	64.5
	2	**42.1**	**65.1**
	5	40.6	63.3

Table 4. ImageNet results while tuning the hyperparameter λ_{div} controlling the importance of our diversity loss. The results are obtained by nearest neighbor models applied on the self-supervised latent space of MAE and CL-MAE based on ViT-B. The top scores are in bold.

Method	N	Acc@1	Acc@5
MAE (baseline) [23]	-	39.2	61.5
CL-MAE	4	42.1	65.0
	5	**42.1**	**65.1**
	6	40.0	63.3

Table 5. ImageNet results while tuning the number of transformer blocks N inside our learnable masking module. The results are obtained by nearest neighbor models applied on the self-supervised latent space of MAE and CL-MAE based on ViT-B. The top scores are in bold.

Method	ViT-B		ViT-L		ViT-H	
	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
MAE [23]	39.2	61.5	44.8	67.0	44.1	66.1
CL-MAE	**42.1**	**65.1**	**45.2**	**67.2**	**45.7**	**68.2**

Table 6. ImageNet results obtained by nearest neighbor models applied on the self-supervised latent space of MAE [23] and CL-MAE (ours) based on various backbones (ViT-B, ViT-L, ViT-H). The top scores for each backbone are in bold.

Number of transformer blocks. Our masking module comprises a configurable number of transformer blocks N. We tune this hyperparameter considering values in the set $\{4, 5, 6\}$ and report the corresponding results in Table 5. The experiments show that choosing $N = 5$ provides the best accuracy rates. Thus, we use $N = 5$ in the next experiments.

4.4. Results

Results on ImageNet. In Table 6, we present the nearest neighbor results for MAE [23] and CL-MAE (ours) using three different backbones for each of the two frameworks, on the ImageNet data set. CL-MAE is consistently better than MAE [23], regardless of the underlying architecture. Our framework brings considerable performance gains for the ViT-B and ViT-H architectures, and moderate gains for ViT-L. According to a battery of paired McNemar's tests, all our gains are statistically significant, at a p-value of 0.001.

Protocol	Method	Aerial Images		Airbus Wind Turbines		Architectural Heritage Elements		Sea Animals		Sport Balls	
		Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5	Acc@1	Acc@5
Nearest Neighbor	MAE (Vit-B) [23]	80.2	94.9	92.1	97.7	76.8	93.4	51.2	76.2	57.6	81.8
	CL-MAE (Vit-B)	**82.6**	**98.1**	**93.4**	**97.9**	**79.3**	**94.0**	**56.2**	**79.7**	**60.2**	**84.5**
	MAE (Vit-L) [23]	83.4	95.7	93.7	98.5	75.6	93.1	52.1	77.0	58.0	82.5
	CL-MAE (Vit-L)	**84.7**	**96.3**	**94.9**	**98.8**	**76.1**	**93.7**	**53.4**	**77.3**	**58.3**	**82.7**
	MAE (Vit-H) [23]	84.0	97.9	94.6	98.4	73.6	92.2	51.1	76.8	**57.7**	81.6
	CL-MAE (Vit-H)	**85.3**	**98.9**	**95.1**	**99.3**	**76.6**	**93.0**	**51.7**	**76.9**	56.1	**82.0**
Linear Probing	MAE (Vit-B) [23]	84.8	97.6	97.8	99.6	84.6	99.6	67.5	92.5	62.9	90.4
	CL-MAE (Vit-B)	**85.4**	**97.9**	**98.8**	**99.9**	**86.8**	**99.9**	**67.9**	**92.7**	**65.8**	**90.7**
	MAE (ViT-L) [23]	84.7	97.7	98.2	99.6	**87.2**	99.7	69.8	93.4	67.4	91.6
	CL-MAE (ViT-L)	**85.9**	**98.1**	**99.1**	**99.9**	**87.2**	**99.8**	**71.0**	**93.8**	**69.0**	**92.1**
	MAE (ViT-H) [23]	86.2	98.5	98.3	99.7	88.3	**99.7**	75.4	95.0	74.6	94.3
	CL-MAE (ViT-H)	**87.1**	**99.3**	**99.3**	**99.9**	**89.6**	99.5	**76.2**	**95.2**	**75.1**	**94.9**

Table 7. Nearest neighbor (top half) and linear probing (bottom half) results on five benchmarks: Aerial Images, Airbus Wind Turbines, Architectural Heritage Elements, Sea Animals, and Sport Balls. The results are reported for MAE [23] and CL-MAE (ours) based on various backbones (ViT-B, ViT-L, ViT-H). The top scores for each backbone on each data set are in bold.

Method	$\mathcal{L}_{\text{Gauss}}$	$\mathcal{L}_{\text{KL}}$	$\mathcal{L}_{\text{div}}$	$\mathcal{L}_{\text{CL}}$	Acc@1	Acc@5
MAE [23]	✗	✗	✗	✗	39.2	61.5
CL-MAE	✓	✗	✗	✗	38.7	61.7
	✓	✓	✗	✗	39.5	61.9
	✓	✓	✓	✗	40.8	62.4
	✓	✓	✓	✓	**42.1**	**65.1**

Table 8. Ablation study on the loss functions used to train our learnable masking module. The results are obtained by nearest neighbor models applied on the self-supervised latent space of MAE and CL-MAE based on ViT-B. The top scores are in bold.

Results on downstream tasks. In Table 7, we present nearest neighbor and linear probing results on Aerial Images [54], Airbus Wind Turbines [17], Architectural Heritage Elements [28], Sea Animals [48], and Sport Balls [10] data sets. The goal of these experiments is to assess the transferability of the self-supervised representations learned by MAE [23] and CL-MAE, using three different backbones (ViT-B, ViT-L and ViT-H). In 57 out of 60 cases, CL-MAE outperforms MAE, with absolute gains varying between $+0.1\%$ and $+4.0\%$. A battery of paired McNemar's tests confirms that our gains are statistically significant, at a p-value of 0.001. For the nearest neighbor protocol, there are 14 out of 30 cases where the absolute gains are higher than 1%. For the linear probing protocol, we have 7 out of 30 cases in which the gains are higher than 1%. The few-shot linear probing experiments (presented in the supplementary) are consistent with the results presented in Table 7. In summary, we conclude that our experiments provide comprehensive evidence indicating that CL-MAE is able to learn superior representations compared to MAE [23].

Ablation study. To assess the individual performance impact of the proposed loss functions, we conduct an ablation study on ImageNet and present the results in Table 8. Note that our learnable masking module needs at least one loss to function. Using solely the Gaussian loss is slightly worse than employing the vanilla MAE. Adding the Kullback-Leibler loss regulates the number of masked patches and improves the model, which becomes marginally better than MAE. The diversity loss plays an important role in further boosting the performance of CL-MAE. The curriculum loss also brings significant performance gains. In summary, the ablation study shows that each and every loss function contributes to the high performance gains of our model.

5. Conclusion

In this paper, we introduced a novel approach for self-supervised representation learning with masked autoencoders, leveraging the concept of curriculum learning. Our method involves generating masks of increasing complexity using a novel learnable masking module. We proposed four losses to ensure that our masking module learns to produce masks that are decisive (close to binary), diverse, and in line with the imposed masking ratio and complexity. Our masking module is jointly trained with MAE, but its reconstruction objective changes during training, from a consensual objective (aiming to help MAE) to an adversarial objective (aiming to confuse MAE), generating an easy-to-hard curriculum learning setup. We conducted comprehensive experiments to compare our framework (CL-MAE) with the vanilla MAE. Our empirical results showed that CL-MAE learns better representations, outperforming the transfer learning capability of MAE. In future work, we aim to extend our analysis to other domains where MAE was successfully employed, *e.g.* video [45] and audio-video [18].

6. Acknowledgments

This work has been funded by Milestone Systems through the Milestone Research Programme at AAU.

References

[1] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Gregoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *arXiv:2304.12210*, 2023. 1

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *Proceeding of ICLR*, 2022. 1, 3

[3] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of ICML*, pages 41–48, 2009. 1, 3

[4] Stefan Braun, Daniel Neil, and Shih-Chii Liu. A curriculum learning method for improved noise robustness in automatic speech recognition. In *Proceedings of EUSIPCO*, pages 548–552, 2017. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of ECCV*, pages 213–229, 2020. 1

[6] Haijia Chen, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. Improving masked autoencoders by learning where to mask. *arXiv:2303.06583*, 2023. 1, 3

[7] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *Proceeding of CVPR*, pages 22742–22751, 2023. 1, 3

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*, pages 1597–1607, 2020. 2, 3

[9] Xinlei Chen and Abhinav Kumar Gupta. Webly supervised learning of convolutional networks. In *Proceeding of ICCV*, pages 1431–1439, 2015. 3

[10] Samuel Cortinhas. Sport Balls - Multiclass Image Classification. https://www.kaggle.com/datasets/samuelcortinhas/sports-balls-multiclass-image-classification, 2022. Accessed: 2023-08-28. 6, 8

[11] Anurag Das, Yongqin Xian, Yang He, Zeynep Akata, and Bernt Schiele. Urban scene semantic segmentation with low-cost coarse annotation. In *Proceedings of WACV*, pages 5978–5987, 2023. 1

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, 2019. 1, 3

[13] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of ICCV*, pages 1422–1430, 2015. 1

[14] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. In *Proceedings of AAAI*, pages 552–560, 2023. 1, 3

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*, 2021. 1, 2, 4, 6

[16] Yueqi Duan, Haidong Zhu, He Wang, Li Yi, Ram Nevatia, and Leonidas J. Guibas. Curriculum DeepSDF. In *Proceedings of ECCV*, pages 51–67, 2020. 3

[17] Airbus DS GEO. Airbus Wind Turbines Patches. https://www.kaggle.com/datasets/airbusgeo/airbus-wind-turbines-patches, 2022. Accessed: 2023-08-28. 6, 8

[18] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual Masked Autoencoders. In *Proceedings of ICCV*, pages 16144–16154, 2023. 8

[19] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Proceedings of ECCV*, pages 540–557, 2021. 1

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of NIPS*, pages 2672–2680, 2014. 2, 4

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent – a new approach to self-supervised learning. In *Proceedings of NeurIPS*, pages 21271–21284, 2020. 2, 3

[22] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *Proceedings of ICML*, pages 2535–2544, 2019. 3

[23] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of CVPR*, pages 16000–16009, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of CVPR*, pages 9729–9738, 2020. 2, 3

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 1, 3

[26] Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Claudiu Popescu, Dim P. Papadopoulos, and Vittorio Ferrari. How Hard Can It Be? Estimating the Difficulty of Visual Search in an Image. In *Proceeding of CVPR*, pages 2157–2166, 2016. 3

[27] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. In *Proceeding of ICML*, page 2304–2313, 2017. 3

[28] Ivan Kobzev and Vasiliev Roman. Architectural Heritage Elements Image Dataset. https://www.kaggle.com/datasets/ikobzev/architectural-heritage-elements-image64-dataset, 2021. Accessed: 2023-08-28. 6, 8

[29] M. Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Proceeding of NeurIPS*, volume 23, 2010. 3

[30] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. SemMAE: Semantic-Guided Masking for Learning Masked Autoencoders. In *Proceedings of NeurIPS*, pages 14290–14302, 2022. 1, 3

[31] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform Masking: Enabling MAE Pre-training for Pyramid-based Vision Transformers with Locality. *arXiv:2205.10063*, 2022. 1, 3

[32] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of CVPR*, pages 9359–9367, 2018. 1

[33] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context Encoders: Feature Learning by Inpainting. In *Proceedings of CVPR*, pages 2536–2544, 2016. 1

[34] Anastasia Pentina, Viktoriia Sharmanska, and Christoph H. Lampert. Curriculum learning of multiple tasks. In *Proceeding of CVPR*, pages 5492–5500, 2014. 3

[35] Pedro O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C. Courville. Unsupervised learning of dense visual representations. In *Proceedings of NeurIPS*, pages 4489–4500, 2020. 2

[36] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom Michael Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of NAACL*, pages 1162–1172, 2019. 3

[37] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. FreeSeg: Unified, Universal and Open-Vocabulary Image Segmentation. In *Proceeding of CVPR*, pages 19446–19455, 2023. 1

[38] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of ICML*, pages 8821–8831, 2021. 3

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2, 6

[40] Yangyang Shi, Martha Larson, and Catholijn M. Jonker. Recurrent neural network language model adaptation with curriculum learning. *Computer Speech & Language*, 33:136–154, 2015. 3

[41] Yuge Shi, N. Siddharth, Philip H.S. Torr, and Adam R. Kosiorek. Adversarial masking for self-supervised learning. In *Proceeding of ICML*, pages 20026–20040, 2022. 1, 3

[42] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of CVPR*, pages 761–769, 2016. 3

[43] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565, 2022. 1, 3

[44] Valentin I. Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Proceedings of NAACL*, pages 751–759, 2010. 3

[45] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Proceedings of NeurIPS*, pages 10078–10093, 2022. 8

[46] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of ICML*, pages 10347–10357, 2021. 1

[47] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Revisiting contrastive methods for unsupervised learning of visual representations. In *Proceedings of NeurIPS*, pages 16238–16250, 2021. 2

[48] Lanz Vencer. Sea animals image dataset. https://www.kaggle.com/datasets/vencerlanz09/sea-animals-image-dataste, 2023. Accessed: 2023-08-28. 6, 8

[49] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of CVPR*, pages 6312–6322, 2023. 1, 3

[50] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M. Alvarez. FreeSOLO: Learning To Segment Objects Without Annotations. In *Proceedings of CVPR*, pages 14176–14186, 2022. 1

[51] Chen Wei, Haoqi Fan, Saining Xie, Chaoxia Wu, Alan Loddon Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceeding of CVPR*, pages 14648–14658, 2022. 1, 3

[52] Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *Proceedings of ICML*, pages 5238–5246, 2018. 3

[53] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. In *Proceedings of ICCV*, pages 22–31, 2021. 1

[54] Gui-Song Xia, Jingwen Hu, Fan Hu, Baoguang Shi, Xiang Bai, Yanfei Zhong, Liangpei Zhang, and Xiaoqiang Lu. AID: A Benchmark Dataset for Performance Evaluation of Aerial Scene Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017. 6, 8

[55] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *Proceedings of CVPR*, pages 9653–9663, 2022. 1, 3

[56] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *Proceedings of NeurIPS*, pages 27127–27139, 2022. 1, 3

[57] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proceedings of ECCV*, pages 649–666, 2016. 1

[58] Minghang Zheng, Peng Gao, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. In *Proceedings of BMVC*, 2020. 1

[59] Liu Zhili, Kai Chen, Jianhua Han, Hong Lanqing, Hang Xu, Zhenguo Li, and James Kwok. Task-customized masked autoencoder via mixture of cluster-conditional experts. In *Proceedings of ICLR*, 2023. 1, 3

[60] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *Proceedings of ICLR*, 2022. 1, 3

[61] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of ICLR*, 2020. 1