Simple Filtering Improves Masked Autoencoders

Takumi Kobayashi[1,2]

[1] National Institute of Advanced Industrial Science and Technology, Japan
[2] University of Tsukuba, Japan
takumi.kobayashi@aist.go.jp

Abstract. Deep models are successfully applied to various fields, while demanding large amount of annotated data for high-performance recognition. To remedy the data-hunger issue, self-supervised learning, especially masked autoencoder (MAE), is a promising approach to effectively pre-train the deep models. The MAE leverages random masking to construct pretext tasks where masked image patches are reconstructed by using unmasked (visible) ones. In vision domain, however, inherent image properties of high redundancy and correlation in local neighbors could interfere with the mask-based pretext tasks. In this study, we analyze two main processes of masking and reconstruction in MAE through the lens of difficulty of the pretext task. The analysis inspires us to propose simple yet effective approaches based on *filtering* to improve random masking as well as raw-pixel reconstruction by properly controlling difficulty of MAE task with a negligible extra computation cost. In the experiments on image classification, the proposed method renders favorable performance improvement to MAE using ViTs.

Keywords: Masked autoencoder · Task difficulty · Filtering

1 Introduction

Self-supervised learning (SSL) attracts keen attention for effectively pretraining large-scale deep models without external supervision. SSL is built upon pretext tasks which are formulated such as by considering knowledge of target domain, e.g., visual image [18, 38, 39, 53]. It can be sophisticated and generalized by leveraging data augmentation techniques in a framework of contrastive learning [3, 5, 7, 14, 20, 21].

On the other hand, masked language modeling (MLM) [2, 11, 29] works effectively for pretraining large-scale language models to exhibit excellent performance on various downstream tasks. The approach of *mask-and-predict* in MLM naturally inspires masked image modeling (MIM) [1, 22, 51] in vision domain by regarding image patches as basic units (tokens) instead of words in NLP. Compared to texts in MLM, however, image signals contain unfavorable characteristics for mask-based pretext tasks where masked (invisible) patches are predicted/reconstructed from the remaining unmasked (visible) patches. Namely, images are continuously represented by RGB-color pixel intensities in contrast to quantized vocabulary of text words, thereby inducing high redundancy in image

patches heavily correlated with neighboring patches [41], which makes it hard to control task difficulty of MIM [1]; in SSL, *not so easy nor difficult* pretext tasks are required for leaning effective image feature representations.

There are various works [24] to address the issues in MIM by improving the input process of *masking* as well as the output process of *patch reconstruction*. While masked autoencoder (MAE) [22] provides data-agnostic random masking in a simple and effective way, the data-adaptive masking [30, 34, 35, 46] can be designed by analyzing visual contents to aggressively mask out object-related patches. On the other hand, in patch reconstruction, the target patch representation can be sophisticated by extracting semantic image features [1, 4, 12, 48]. Those approaches address the MIM task difficulty by exploiting semantic image characteristics which contribute to performance improvement at the sake of computation efficiency; they resort to additional networks and/or training processes that increase computation cost, interfering with simplicity and efficiency of MIM approach [22] which are useful for enhancing model scalability in SSL.

In this paper, we propose a simple and efficient approach to effectively control the *task difficulty* of MIM while keeping computation efficiency of MAE [22]. The proposed method is built upon *filtering*, an efficient data-agnostic technique, which plays different roles in the two key processes of masking and patch reconstruction; it works for increasing difficulty in masking while properly easing the reconstruction task. While the task difficulty in the masking is related to semantic image content as mentioned above, even the data-agnostic mask shapes are potentially involved with the task difficulty [22] due to high correlation among neighboring patches that images intrinsically possess. We first attempt to quantitatively measure task difficulty derived from a mask shape by means of mean nearest neighbor distance [8] over the mask pattern. Then based on the measurement, we propose a *pooling*-based approach to produce masks of favorable task difficulty. On the other hand, image pixels are vulnerable to noises and perturbations irrelevant to intrinsic image characteristics, making it harder to reconstruct patch appearance. We analyze the reconstruction process through the lens of image frequencies to propose a simple image *smoothing* for favorably easing the patch reconstruction. Thus, the proposed method solely resorts to simple filtering which consumes only negligible extra computation cost, thereby keeping both simplicity and efficiency of MAE [22].

Our contributions are summarized as follows:

- In a masking process, we measure task difficulty based on nearest neighbor distance between masked and unmasked patches. Based on the measurement, we propose a pooling-based approach to endow random masking with capability of controlling the task difficulty.
- Through the analysis of patch reconstruction process, we clarify a biased representation of the predicted patches from a viewpoint of image frequencies. To cope with the bias, we propose an approach to efficiently transform the target patch representation by image smoothing.

– In the experiments, the proposed filtering-based MAE is thoroughly analyzed from various aspects, demonstrating that our method effectively boosts performance of MAE on image recognition tasks.

1.1 Related works

Masked image modeling. Witnessing the success of masked language modeling (MLM) [2, 11, 29] in NLP fields, masked image modeing (MIM) [1, 22] is also proposed for self-supervised learning of vision models. Masked autoencoders (MAE) [22] formulates the MIM by randomly masking image patches and reconstructing them from remaining unmasked ones, which exhibits high synergy with vision transformers (ViTs) [13]; only visible (unmasked) patches are fed into ViT to reduce computation cost significantly. There are some works [27, 52] to theoretically analyze MAE. We focus on this MAE framework due to its simplicity and computation efficiency.

In comparison to words which are quantized codes of high-level semantics, the distinctive image modality of continuous and highly redundant pixel representation could impede the self-supervised learning, especially in terms of MIM task difficulty; neither too easy nor too difficult task is preferable for the self-supervised learning. MAE addresses the issues by using high masking ratio and lightweight decoder to reconstruct patches, and the processes of masking and reconstruction can be further enhanced as follows.

Masking. Task difficulty in masking can be related to (*i*) *what to mask* and (*ii*) *how to mask*. The former one is addressed by content-based masking which is designed such as by using attention map of teacher network [34], group-clustering of attention weights [17], learning-based task difficulty control [35,43], part-based semantic segmentation [30] and estimation of patch reconstruction loss [46]; they demand extra computation cost to extract content-related image characteristics.

The latter masking strategy (*ii*) is constructed in a content-agnostic approach which resorts to random selection of naive patches [22], structured and naive block-based superset of the patches [1,25,51]. In particular, random masking [22] plays a pivotal role in MIM as it performs in a computationally efficient way on the assumption that meaningful image contents are statistically covered due to the randomness; the above-mentioned content-based masking (*i*) also employs such random masking to effectively tune task difficulty. Thus, we focus on the content-agnostic random masking. In the literature, ColorMAE [23] improves performance of MAE just by devising the content-agnostic masking strategy, which provides a rationale that the masking strategy (*ii*) contributes to essential task difficulty of MAE. The ColorMAE is related to ours as it sophisticates the random masking by means of band-pass filtering from image processing viewpoint in a heuristic manner. In contrast, we formulate the task difficulty induced by mask shapes to quantitatively characterize content-agnostic masking and accordingly propose pooling-based masking—a simpler and more efficient approach—in a principled way.

Patch reconstruction. Beyond raw-pixel representation [22,51], more semantic characteristics are leveraged to describe the reconstruction target patch such as by using color clustering [4], image features [48] via HOG [9] and visual tokenizer with extra models [1, 12]. They could ease the task difficulty by suppressing information irrelevant to intrinsic visual characteristics, though being accompanied by computation burden. For efficient MAE, we focus on raw-pixel reconstruction [22] and analyze the predicted patch appearance from a frequency viewpoint, which leads to our simple approach that manipulates target patches via image smoothing to reduce task difficulty efficiently. Our analysis based on image frequency is different from the approaches to alter MAE by utilizing frequency-based image representation for masking targets [50, 55] and decoders [31]. While high-frequency components are cut off to smooth a target image in [32], we simply incorporate filtering into MAE [22] which softly suppresses high-frequency components, while retaining the simplicity and efficacy without changing the architecture of MAE; it is noteworthy that our smoothing process is combined with mask filtering (pooling) to consistently work on task difficulty (Sec. A in the supplementary material).

2 Method

We address task difficulty of MAE in the two key processes of *random masking* and *raw-pixel patch reconstruction*.

2.1 Masking

As discussed in Sec. 1.1, we address the task difficulty from the aspect of *content-agnostic* masking, as implied in [23]; it could be combined with content-based difficulty as our future work.

Preliminary. Suppose an image is spatially divided into $H \times W$ patches; the whole image is described by $\mathcal{I}=\{\boldsymbol{x}_i\}_{i=1}^{n}$ where $n = HW$ and each $s_h \times s_w$-sized patch $\boldsymbol{x}_i \in \mathbb{R}^{s_h \times s_w \times 3}$ is associated with a 2-D position vector $\boldsymbol{z}_i \in [1,\cdots,H] \times [1,\cdots,W]$. The masking process separates $\mathcal{I}$ into two subsets of masked (invisible) and unmasked (visible) patches as $\mathcal{I} = \mathcal{M} \cup \bar{\mathcal{M}}$ where $\mathcal{M} = \{\boldsymbol{x}_1^{\texttt{mask}}, \cdots, \boldsymbol{x}_m^{\texttt{mask}}\}$ and $\bar{\mathcal{M}} = \{\boldsymbol{x}_1^{\texttt{unmask}}, \cdots, \boldsymbol{x}_{n-m}^{\texttt{unmask}}\} = \mathcal{I} \backslash \mathcal{M}$; in MAE literature, $p = \frac{m}{n}$ is referred to as a mask ratio.

Task difficulty. As the pretext task in MAE is to reconstruct (predict) $\mathcal{M}$ from $\bar{\mathcal{M}}$, difficulty of the MAE task is connected to *correlation* between patches of $\boldsymbol{x}_i^{\texttt{unmask}}$ and $\boldsymbol{x}_j^{\texttt{mask}}$. Considering redundancy and continuity of image patterns, prediction of *near-by* patches would be an easy task; high mask ratio p is practically employed, say $p = 0.75$ [22], so as to possibly put $\boldsymbol{x}^{\texttt{mask}}$ away from $\boldsymbol{x}^{\texttt{unmask}}$ for constructing moderately difficult tasks. Thus, the task difficulty can be involved with spatial adjacency between $\mathcal{M}$ and $\bar{\mathcal{M}}$ in a content-agnostic manner.

We measure the adjacency between those patch sets by using *mean nearest neighbor distance* as follows.

$$\mathtt{d}(\mathcal{M}) = \mathop{\mathrm{E}}_{\boldsymbol{z}^{\texttt{mask}} \in \mathcal{M}} \min_{\boldsymbol{z}^{\texttt{unmask}} \in \bar{\mathcal{M}}} \|\boldsymbol{z}^{\texttt{mask}} - \boldsymbol{z}^{\texttt{unmask}}\|_2^2, \tag{1}$$

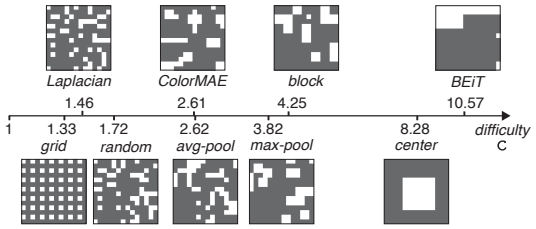


Fig. 1: Difficulty scores $\mathtt{c}(\mathbb{M}_p)$ in (2) for various types of masking strategies $\mathbb{M}_p$. The scores are empirically computed by repeatedly sampling masks of 14×14 with $p = 0.75$. Gray/white points indicate *masked*/*unmasked* patches.

where $\boldsymbol{z}$ indicates a 2-D patch position, and the unmasked set $\bar{\mathcal{M}}$ is deterministically given by $\bar{\mathcal{M}} = \mathcal{I}\backslash\mathcal{M}$. In (1), we compute *nearest-neighbor distance* at each masked patch $\boldsymbol{z}^{\mathtt{mask}}$ and then average the distance over the mask set $\mathcal{M}$ as in mean nearest neighbor distance mainly addressed in material science [8,15,16,44]. Thereby, the smaller distance score (1) indicates that masked patches $\mathcal{M}$ tightly adjoin the unmasked ones $\bar{\mathcal{M}}$, leading to an easier MAE task. Thus, the score (1) can reflect the task difficulty.

By using the distance score (1), we formulate the task difficulty induced by a masking strategy $\mathbb{M}_p$ that produces a mask $\mathcal{M}$ in a (random) manner with a mask ratio p;

$$\mathtt{c}(\mathbb{M}_p) = \mathop{\mathrm{E}}_{\mathcal{M}\sim\mathbb{M}_p} \mathtt{d}(\mathcal{M}). \tag{2}$$

Analysis of task difficulty $\mathtt{c}$**.** Fig. 1 depicts various types of masking strategies $\mathbb{M}_p$ and their scores $\mathtt{c}(\mathbb{M}_p)$ with $p = 0.75$. *Grid* masking that uniformly locates unmasked patches in a grid manner is regarded as an easier task of lower $\mathtt{c}$ while *center* and *BEiT* approach [1] render much more difficult tasks of higher $\mathtt{c}$. It should be noted that these extreme masking poorly work in MAE as shown in Table 1, which thus inspires us to design masking of *moderate* difficulty score. The *random* masking, a popular approach in MAE literature, produces rather moderate difficulty, compared to *grid*. The following analytical scores $\mathtt{c}$ of *grid* and *random* masking facilitate further analysis; the proofs are shown in Sec. B of the supplementary material.

Proposition 1 *The grid masking* $\mathbb{M}_p^{\mathtt{grid}}$ *with* $p = 0.75$ *produces the score of* $\mathtt{c}(\mathbb{M}_{p=0.75}^{\mathtt{grid}}) = \frac{4}{3}$.

Proposition 2 *The random masking* $\mathbb{M}_p^{\mathtt{rand}}$ *provides*

$$\mathtt{c}(\mathbb{M}_p^{\mathtt{rand}}) \approx (1-p^4) + 2p^4(1-p^4) + 4p^8(1-p^4) + O(p^{12}). \tag{3}$$

Specifically, in case of $p = 0.75$, $\mathtt{c}(\mathbb{M}_{0.75}^{\mathtt{rand}}) \approx 1.55$[3].

[3] The empirical score (1.72 in Fig. 1) is practically deviated from the theoretical one (1.55) due to boundary effect, as shown in Sec. B.

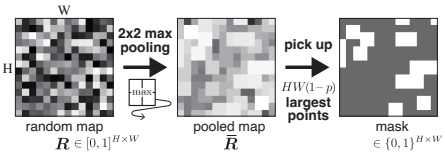


Fig. 2: Our max-pooled masking. Random-value matrix $\boldsymbol{R} \in [0,1]^{H\times W}$ is first filtered by max-pooling, followed by selecting the top $HW(1-p)$ points for *unmasked* patches.

These propositions show that the *random* masking intrinsically leans toward *grid* masking which produces the easier MAE task. Thus, we explore masking strategies of higher difficulty $\mathtt{c}$ than the *random* masking.

Pooling-based masking. The score (1) is rewritten to

$$\mathtt{d}(\mathcal{M}) = \mathop{\mathrm{E}}_{\boldsymbol{z}^{\mathtt{mask}}\in\mathcal{M}} \sum_{\boldsymbol{z}\in\mathcal{I}} \mathtt{p}(\boldsymbol{z}=\mathtt{nn}(\boldsymbol{z}^{\mathtt{mask}};\bar{\mathcal{M}})|\boldsymbol{z}^{\mathtt{mask}})\|\boldsymbol{z}^{\mathtt{mask}}-\boldsymbol{z}\|_2^2, \tag{4}$$

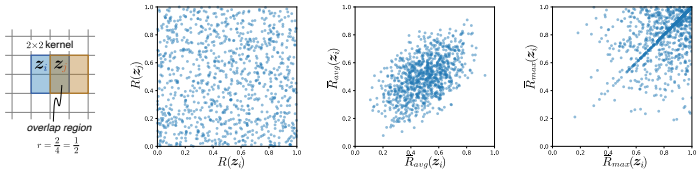
where $\mathtt{p}$ is a probability function to encode the relationship between the masked patch $\boldsymbol{z}^{\mathtt{mask}}$ and its nearest neighbor in $\bar{\mathcal{M}}$, $\mathtt{nn}(\boldsymbol{z}^{\mathtt{mask}};\bar{\mathcal{M}}) = \arg\min_{\boldsymbol{z}^{\mathtt{unmask}}\in\bar{\mathcal{M}}} \|\boldsymbol{z}^{\mathtt{mask}} - \boldsymbol{z}^{\mathtt{unmask}}\|_2$. The probability function $\mathtt{p}$ is characterized by the masking strategy $\mathbb{M}_p$ and plays a crucial role for the difficulty score $\mathtt{c}(\mathbb{M}_p)$. As shown in Fig. 1, distributions of unmasked patches are related to the difficulty scores; the scores shift from low (easy) to high (difficult) as the unmask distribution becomes localized, affecting $\mathtt{p}$. In *random* masking, the probability of masking at each patch $\boldsymbol{z} \in \mathcal{I}$ is independent in *i.i.d.* fashion, therefore embedding *no* correlation between a pair of patches to provide less localized unmask distribution. From a probabilistic viewpoint, the pair-wise relationships could be modeled by *Copulas* [37]. It, however, is hard to flexibly model the copula on a spatial 2D space [40] in a computationally efficient way. Thus, we resort to tweaking *i.i.d.* random-value 2D map which are practically utilized to construct *random* masking.

In *random* masking, uniform random numbers are assigned to respective patch positions $\boldsymbol{z} \in \mathcal{I}$ to form a random-value map $\boldsymbol{R} \in [0,1]^{H\times W}$, like a random noise map. Then, patch positions $\boldsymbol{z}_i$ of $n(1-p)$ largest values $R(\boldsymbol{z}_i)$ are picked up as unmasked ones. As those random numbers are drawn in an *i.i.d* manner, there is no correlation among positions (Fig. 3*b*). In this study, to control the task difficulty, i.e., mean nearest neighbor distance (4), we embed correlation into near-by patch positions through *filtering* on the random map $\boldsymbol{R}$; specifically, we apply *average/max-pooling* to $\boldsymbol{R}$.

In the case of average pooling with $k \times k$ kernel, two near-by patches exhibit the following correlation;

Proposition 3 *Let $\bar{\boldsymbol{R}}$ be an avg-pooled map of $\boldsymbol{R}$ with $k \times k$ kernel, and $\bar{R}(\boldsymbol{z})$ be an element of $\bar{\boldsymbol{R}}$ at the patch position $\boldsymbol{z}$. Joint probability for two patches $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ is given by*

$$\mathtt{p}(\bar{R}(\boldsymbol{z}_i), \bar{R}(\boldsymbol{z}_j)) \approx \mathcal{N}\left(\left[\frac{1}{2}, \frac{1}{2}\right], \frac{1}{12k^2}\begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right), \tag{5}$$



(*a*) kernel (*b*) raw random (*c*) avg-pool (*d*) max-pool

Fig. 3: Relationship of random values at two adjacent points. (*a*) Two kernels are overlapped at near-by points $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$. (*b*,*c*,*d*) Distribution of pair-wise values at the near-by points on raw ($\boldsymbol{R}$), 2×2 avg-pooled ($\bar{\boldsymbol{R}}_{avg}$) and max-pooled ($\bar{\boldsymbol{R}}_{max}$) random-value map through 1000-times sampling.

where $r \in [0, 1)$ *is a overlap ratio between two kernel receptive fields at* $\boldsymbol{z}_i$ *and* $\boldsymbol{z}_j$*, as shown in Fig. 3a.*

By applying average-pooling, random values at two patches are correlated due to the overlap between their kernel receptive fields, as demonstrated in Fig. 3*c*. Thus, if a patch is picked up as unmasked, its neighboring patches would be likely selected as well. Similarly, max-pooling also contributes to enhancing correlation between near-by patches, as empirically shown in Fig. 3*d*; the max operation further encourages the likelihood to jointly select near-by patches as unmasked ones. Through the pooling on $\boldsymbol{R}$, spatial distribution of unmasked patches are localized, increasing the difficulty score $\mathtt{c}$ as shown in Fig. 4; larger kernel size $k \times k$ enhances correlation among patches to increase the difficulty. On the contrary, contrast-enhancing *Laplacian* filtering on $\boldsymbol{R}$ reduces correlation to decrease $\mathtt{c}$, approaching $\mathtt{c} = 1.33$ of *grid* masking, as shown in Fig. 1.

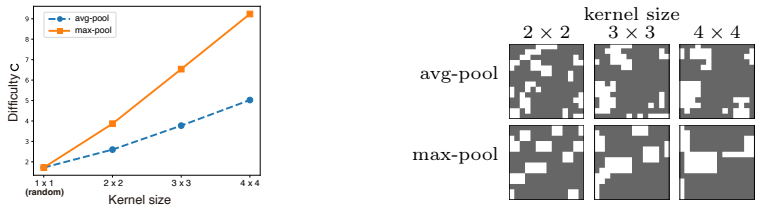
Fig. 2 summarizes our masking process that first draws a random map $\boldsymbol{R} \in [0, 1]^{H \times W}$ as in *random* masking and then filters it via *avg/max-pool* with $k \times k$ kernel, followed by finally picking up $n(1 - p)$ unmasked points to construct a mask pattern. The task difficulty $\mathtt{c}$ can be controlled by the kernel size $k \times k$ and type of pooling, average or max.

2.2 Reconstruction

MAE task is to reconstruct masked patches. Though it is akin to denoising image autoencoders in an image enhancement literature, the primary goal of MAE is to learn effective feature representation of a backbone (encoder) ViT rather than to precisely reconstruct the image patch appearance in detail. Thus, the image reconstruction in MAE is just a means to access intrinsic image content [27]. The reconstruction for raw pixels efficiently works in MAE [22, 51] which can be analyzed in detail as follows.

Frequency analysis. We analyze what kind of characteristics in image patches are recovered by the MAE, through the lens of image *frequency* based on discrete cosine transform (DCT). We compute the following frequency-wise normalized error between a masked patch image $\boldsymbol{x}^{\mathtt{mask}}$ and a predicted patch $\hat{\boldsymbol{x}}^{\mathtt{mask}}$;

$$\mathtt{e}_f(\boldsymbol{x}^{\mathtt{mask}}, \hat{\boldsymbol{x}}^{\mathtt{mask}}) = \frac{1}{2} \mathop{\mathrm{E}}_{f_h, f_w | f_h + f_w = f} \frac{|\mathtt{DCT}_{f_h, f_w}(\boldsymbol{x}^{\mathtt{mask}}) - \mathtt{DCT}_{f_h, f_w}(\hat{\boldsymbol{x}}^{\mathtt{mask}})|^2}{|\mathtt{DCT}_{f_h, f_w}(\boldsymbol{x}^{\mathtt{mask}})|^2 + |\mathtt{DCT}_{f_h, f_w}(\hat{\boldsymbol{x}}^{\mathtt{mask}})|^2}, \quad (6)$$



(*a*) Difficulty scores (*b*) avg/max-pooled masks

Fig. 4: Pooling-based masking with various-sized kernels.

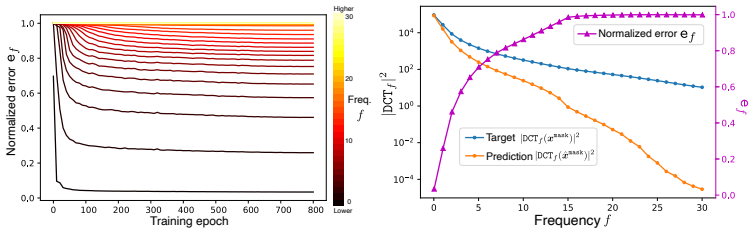
where $f_h \in \{0, \cdots, s_h - 1\}$ and $f_w \in \{0, \cdots, s_w - 1\}$ are vertical and horizontal frequencies of 2D-DCT for an image patch $\boldsymbol{x} \in \mathbb{R}^{s_h \times s_w \times 3}$, and $f = f_h + f_w \in \{0, \cdots, s_h + s_w - 2\}$ indicates a total frequency; the normalized error (6) is bounded in $0 \leq \mathsf{e}_f \leq 1$.

Fig. 5*a* reports the error scores during the training of ViT-S [13] on ImageNet-100 [47] with $s_h = s_w = 16$ on an image of 224×224. It clearly shows that the lower-frequency components are recovered at earlier training epochs and then the higher ones are progressively dealt with as the training proceeds. It is noteworthy that most higher frequencies are *not* reconstructed throughout the training; roughly speaking, top half of frequencies are untouched in this MAE reconstruction, as shown in Fig. 5*b* demonstrating that the predicted image patch contains weak amplitudes at higher frequencies. Thus, we can say that the reconstruction in MAE is biased toward *lower*-frequency components and hardly improves the *higher*-frequency ones. The higher-frequency image components are vulnerable to noise and/or textures less relevant to the intrinsic image contents, which makes the reconstruction task harder. The increased task difficulty would impede MAE learning; minimizing the discrepancy in higher-frequency components might unfavorably encourage the model to focus on superficial image appearance which is problematic in MAE-based feature representation learning [27].

Target image smoothing. Based on the above analysis regarding image frequency, we remove higher-frequency components from the target image (patch) to facilitate feature learning in MAE. To that end, we can *again* resort to *filtering* for smoothing the target patches; in this study, Gaussian filter is applicable as a standard low-pass filter. It should be noted that we inject blur degradation to *target* images via smoothing, in an opposite way to denoising autoencoders [45, 49] which applies such degradation to *input* images. Our MAE loss function is thus defined as mean squared error between *smoothed* target patches $\mathsf{g}_\sigma * \boldsymbol{x}^{\mathtt{mask}}$ and predicted patches $\hat{\boldsymbol{x}}^{\mathtt{mask}}$;

$$\ell_{\mathrm{sm}}(\boldsymbol{x}^{\mathtt{mask}}, \hat{\boldsymbol{x}}^{\mathtt{mask}}) \propto \|\mathsf{g}_\sigma * \boldsymbol{x}^{\mathtt{mask}} - \hat{\boldsymbol{x}}^{\mathtt{mask}}\|_2^2, \tag{7}$$

where g_σ indicates a Gaussian filter with a scale σ. This smoothing filter effectively reduces task difficulty of MAE, promoting feature learning, in a negligible computation cost.



(a) Error dynamics (b) Amplitude

Fig. 5: Normalized frequency-wise errors (6) on ImageNet-100 by using ViT-S. (a) Error dynamics along training epochs, and (b) errors and amplitudes of target and predicted patches across frequencies.

3 Results

The proposed MAE, which is equipped with two types of *filtering* in the process of masking and reconstruction, is applied to pretraining ViTs and analyzed in terms of empirical performance on image classification tasks.

Implementation. ViT models [13] are pretrained on the respective datasets by Adam-W optimizer [33] with a batch size of 1024 in the training recipe given in [22] over 2000 epochs for Cifar-10 dataset and 800 epochs for ImageNet datasets. Following a practice in MAE literature, we set a mask ratio $p = 0.75$. Performance is evaluated by image classification accuracies (%) measured in a protocol of such as linear probe [22] on a validation set of a dataset. Details are shown in Sec. D of the supplementary material.

3.1 Analysis of proposed filtering processes

Filtered masking. Filtering random-value map $\boldsymbol{R}$ is capable of controlling the task difficulty (2) in the masking process. Various types of masking strategies, exhibiting variety of task difficulties, are compared in Table 1; examples of those masks are illustrated in Fig. 1. In fixed masking, *grid* and *center* render low and high difficulties, respectively, both of which work poorly since they produce too easy and difficult tasks. The structured masking including *block* [25, 51] and *BEiT* [1] reduces randomness or variety in terms of masked patch positions, thereby degrading performance especially on Cifar dataset which lacks diversity of image appearance.

On the other hand, filtered masking retains such a variety of mask positions as in *random* masking. Thus, *Laplacian* masking produces superior performance to *grid* masking though they exhibit similar (low) difficulty. The *Laplacian*, however, reduces correlation among near-by patches on $\boldsymbol{R}$, leading to lower difficulty as well as degrading performance in comparison to *random* masking. Our *avg/max-pool* masking enhances the neighbor correlation to favorably increase task difficulty, improving performance. The performance results in Table 1 show that the task difficulty of $\mathtt{c} = 3 \sim 4$ seems to work well and too large-sized kernel degrades performance by increasing difficulty too much; especially, *max-pool* with 4×4 kernel provides *too difficult* task, thus deteriorating performance.

Table 1: Classification accuracies (%) by various types of masking on Cifar-10 [28] and ImageNet-100 [47] datasets.

Masking	Type	Difficulty c	Cifar-10	ImageNet-100
random	-	1.72	79.18±0.02	68.99±0.11
grid	fix	1.33	56.17±0.04	58.82±0.12
center	fix	8.28	45.99±0.04	60.65±0.05
BEiT	structure	10.57	65.46±0.02	68.03±0.04
block (2×2)	structure	4.25	77.87±0.01	71.05±0.11
ColorMAE	band-pass	2.61	79.64±0.01	71.46±0.11
Laplacian	3×3 filter	1.46	77.41±0.03	66.68±0.15
avg-pool	2×2 filter	2.62	81.27±0.01	71.25±0.01
avg-pool	3×3 filter	3.75	80.28±0.03	71.23±0.03
avg-pool	4×4 filter	5.04	80.67±0.02	70.94±0.14
max-pool	2×2 filter	3.83	**82.16**±0.03	**72.07**±0.09
max-pool	3×3 filter	6.55	80.43±0.01	71.45±0.11
max-pool	4×4 filter	9.18	78.46±0.01	69.93±0.14

Table 2: Classification accuracy (%) on ImageNet-100 with various smoothing approaches in the reconstruction process.

(a) Smoothing filter

σ	*random*	*max-pool*
(0)	68.99±0.11	72.07±0.09
1	70.19±0.08	72.37±0.10
2	71.34±0.04	73.05±0.09
4	**72.10**±0.06	**73.29**±0.03
8	71.56±0.03	72.07±0.07
Sharp	58.94±0.04	64.51±0.04
High-freq. cut [32]	69.34±0.09	71.26±0.03

(b) Down-scaling

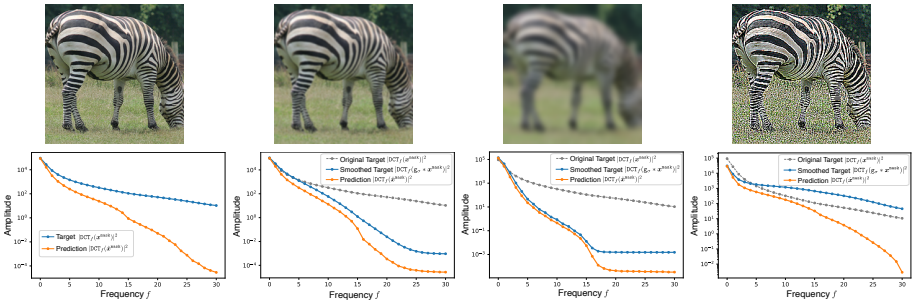
Down-scale factor	*random*	*max-pool*
1	68.99±0.11	72.07±0.09
2	70.18±0.17	71.79±0.06
4	**71.27**±0.08	**72.45**±0.07
8	70.69±0.03	70.65±0.07

Based on the analysis, we conjecture that *max-pool* with 2×2 kernel is favorably compatible for MAE.

The proposed task difficulty c similarly characterizes the *ColorMAE* masking [23] which is derived from an image processing perspective based on band-pass filtering. *ColorMAE* exhibits $\mathtt{c} = 2.61$, larger than $\mathtt{c} = 1.72$ of random masking, to improve performance. Specifically, the performance is close to that of our 2×2 *avg-pool* masking as their difficulty scores are similar. Thus, it is possible to validate the *ColorMAE* masking through our task difficulty c, while *ColorMAE* is inferior to our 2×2 *max-pool* masking.

Filtered reconstruction target. We then analyze the smoothing process (7) on a target image from a *frequency* viewpoint; we apply it with *random* and 2×2 *max-pool* masking on ImageNet-100 dataset by using ViT-S.

• **Smoothing scale σ:** A target image is smoothed by applying Gaussian filter equipped with a scale parameter σ. The performance comparison across various σ is shown in Table 2a while their effects on images are analyzed in Fig. 6. The sim-



(*a*) Original (*b*) $\sigma = 1$ (*c*) $\sigma = 4$ (*d*) Sharp

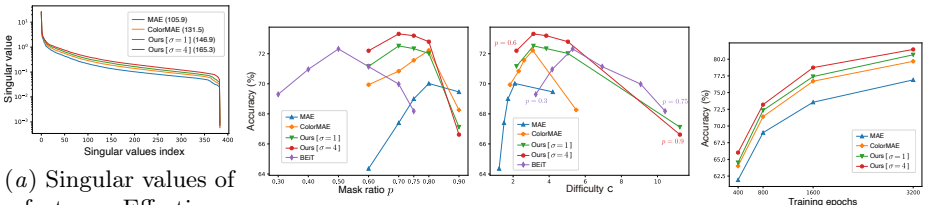
Fig. 6: Examples of smoothed images (top) with their frequency amplitudes (bottom).

ple smoothing filter improves performance effectively on both *random* and 2×2 *max-pool* masking. It is noteworthy that even subtle smoothing of $\sigma = 1$ (Fig. 6*b*) enhances MAE and the performance is maximized by $\sigma = 4$ which smooths out images (Fig. 6*c*). As shown in bottom row of Fig. 6, the smoothing filter reduces higher-frequency components in target images, which facilitates reconstruction by focusing MAE on lower-frequency domains. In other words, as the smoothed images (Fig. 6*bc*) still render semantic-level recognition regarding objects, MAE could encode more intrinsic image characteristics rather than superficial image appearance. To further validate the efficacy of smoothing, we compare it with unsharp-masking filter [19] that enhances image contrast (sharpness). It lifts up higher-frequency bands (Fig. 6*d*), making it harder to reconstruct patches via MAE and thereby significantly degrading performance as shown in Table 2*a*. Therefore, these results demonstrate that the higher-frequency image components are less relevant to the primary image characteristics embedded in this ImageNet-100 dataset and even simple image smoothing improves MAE by favorably reducing task difficulty. Our smoothing approach is also compared with the method that cuts off higher-frequency components [32]. Our filter-based approach softly suppresses high-frequency components via weights (Sec. C in the supplementary material) without cutting them off, thus being effective to improve performance.

• **Down-scaling:** Other than image smoothing, higher-frequency components can also be reduced by down-scaling (via bilinear interpolation) for target patches. The down-scaling apparently reduces the reconstruction task complexity as the dimensionality, i.e., spatial size, of the target is decreased; for example, 16×16 patch is converted into 8×8 by a down-scaling factor of 2. The performance results of down-scaling are shown in Table 2*b*. Due to suppressed high-frequency components via down-scaling, the performance is slightly improved; the best performance is produced by the down-scaling factor of 4 roughly in accordance with Gaussian filter of $\sigma = 4$. It is, however, inferior to image smoothing (Table 2*a*). The downscaling retains some higher-frequency components, known as anti-aliasing issues, which would impede the MAE training.

Table 3: Classification accuracy (%) on ImageNet-100 in comparison to frequency-based losses.

Loss	*random*	*max-pool*
Focal Freq. loss [26]	68.55±0.11	70.94±0.10
Freq. L_1 loss [50]	63.17±0.08	66.03±0.12
L_2 loss in MAE [22]	68.99±0.11	72.07±0.09
Ours (7) via image smoothing $\sigma = 1$	70.19±0.08	72.37±0.10
Ours (7) via image smoothing $\sigma = 4$	**72.10**±0.06	**73.29**±0.03



(*a*) Singular values of features. Effective ranks [42] are shown in parentheses.

(*b*) Mask ratio p. Various p lead to diverse difficulty `c`.

(*c*) Training epoch

Fig. 7: Performance analysis of our MAE on ImageNet-100 using ViT-S.

• **Frequency loss:** The frequency-based perspective can be embedded into FFT-based loss functions [26, 50] which could be compared with our loss based on image smoothing (7). The comparison results are shown in Table 3, demonstrating superiority of our simple approach; this analysis is detailed in Sec. C of the supplementary material. It is noteworthy that our method (7) is computationally efficient just by smoothing an image without manipulating structures of loss.

3.2 Performance analysis

We then analyze performance of our MAE using 2×2-*max-pool* masking and image smoothing (7), in comparison to original MAE [22] using *random* masking and ColorMAE [23].

Feature representation. Fig. 7*a* shows singular values of features with the effective rank [42], demonstrating diversity of the learned feature representation. The proposed method enhances uniformity of singular values as well as effectively enlarges the rank of features, which is favorable in terms of robustness and generalization [36, 52]. In terms of masking, the task difficulty is enhanced via *max-pool* to let MAE look into the intrinsic image characteristics beyond a superficial patch correlation between masked and unmasked ones. The intrinsic image representation is further enhanced by smoothing out irrelevant high-frequency components from target images in patch reconstruction. These two different types of filtering work for promoting feature representation learning in our MAE.

Table 4: Performance results by U-MAE [52] and D-MAE [49] on ImageNet-100 using ViT-S. In U-MAE, we report classification accuracies (%) while showing certified accuracy (%) at various radius r for D-MAE [49].

			Ours	
Method	Orig.	ColorMAE	$\sigma = 1$	$\sigma = 4$
U-MAE	69.36±0.05	67.96±0.11	73.31±0.11	**74.14**±0.10
D-MAE				
$r = 0$	57.20±0.28	57.27±0.09	61.80±0.16	**62.27**±0.19
$r = 0.25$	51.40±0.16	52.93±0.25	56.93±0.25	**57.13**±0.74
$r = 0.5$	45.53±0.34	46.87±0.34	48.40±0.65	**52.60**±0.28
$r = 0.75$	38.20±0.28	38.73±0.34	39.53±0.19	**43.00**±0.16

Mask ratio. The mask ratio p ia a key hyper-parameter in MAE, while $p = 0.75$ is a standard practice in image-based MAE [22]. Fig. 7*b* shows performances at various mask ratio p. We also apply *BEiT* masking which works on lower p [1]; it prefers $p = 0.5$, though increasing computation cost due to larger number of visible (unmasked) patches. By controlling the task difficulty, our MAE is robust against various $p < 0.9$; performance gap between $p = 0.75$ and $p = 0.6$ is about 1% in our MAE which is smaller than 5% by *random* masking in original MAE and 2% by ColorMAE. On the other hand, the performance is degraded at $p = 0.9$, *as expected*, since the task difficulty becomes too high $\mathsf{c} = 11.24$ at $p = 0.9$. Roughly speaking, through the lens of task difficulty c, the masking strategies exhibit rather consistent performance tendency that the performance is optimal around $\mathsf{c} = 4$, while deteriorating at extreme difficulties such as $\mathsf{c} < 2$ and $\mathsf{c} > 8$; on the other hand, the performances exhibit diverse behavior across mask ratio p. Thus, it is noteworthy that the task difficulty c in (2) would be roughly a useful measure to design p, such as by excluding bad masking of extreme (too low or too high) difficulty to avoid performance degradation.

Training epoch. Due to enhanced task difficulty, our MAE approach also enjoys the larger number of training epochs for increasing performance as shown in Fig. 7*c* in accordance with other MAEs.

Application to variants of MAE. As our method touches only two processes of masking and patch reconstruction in MAE, it is straightforwardly applicable to the other variants of MAE. We incorporate the proposed method into U-MAE [52] and D-MAE [49] which address feature uniformity and robustness, respectively. The performance results are shown in Table 4, demonstrating favorable improvement by our method.

Evaluation on ImageNet-1K. The proposed MAE is evaluated on ImageNet-1K [10] by using ViT-B [13]. The performance results are shown in Table 5. The proposed method effectively boosts performance of MAE on two evaluation metrics of linear probe and fine-tuning, while keeping simplicity and computation efficiency of MAE [22]; our method also retains the GPU memory consumption while ColorMAE [23] demands additional memory to store pre-computed noise

Table 5: Performance results on ImageNet-1K using ViT-B; classification accuracy (%) on linear probe and fine-tuning. † and ‡ mean the results borrowed from respective papers and [6] while ∗ indicates our implementation.

Method	Epoch	Lin-probe	Fine-tune
non MIM			
MoCo v3 [7]‡	600	76.2	83.0
DINO [3]‡	1600	77.3	83.3
MAE variants			
U-MAE [52]†	200	58.5	83.0
CAE [6]‡	800	68.6	83.8
CAE [6]‡	1600	71.4	83.9
MAE with data-adaptive masking			
SemMAE [30]†	800	65.0	83.34
HPM [46]†	800	-	84.2
MAE with data-agnostic masking			
BEiT [1]†	800	56.7	83.2
ColorMAE [23]†	800	68.67	83.6
MAE [22]*	800	65.14	83.44
Ours ($\sigma = 1$)	800	70.07	**83.75**
Ours ($\sigma = 4$)	800	**70.42**	83.64
ColorMAE [23]†	1600	70.85	83.8
MAE [22]*	1600	66.71	83.65
Ours ($\sigma = 1$)	1600	72.26	**84.03**
Ours ($\sigma = 4$)	1600	**72.55**	83.78

map. While our smoothing with $\sigma = 4$ works well for linear probing, that of $\sigma = 1$ provides further boost in a fine-tuning scenario; weak smoothing by $\sigma = 1$ enables the model to extract subtle yet effective features which could be enhanced by further fine-tuning on ImageNet-1K.

The ViT-B models pre-trained on ImageNet-1K are then applied to a downstream task of semantic segmentation on ADE-20K dataset [56] in a framework of SETR [54], producing favorable performance improvement as shown in Table 6.

4 Conclusion

We have proposed simple yet effective approaches to enhance MAE. Through the analyses of two main processes of masking and reconstruction in MAE from a viewpoint of task difficulty, our methods are formulated by simple *filtering* to control the task difficulty; *max-pooled* masking enhances the difficulty while *image smoothing* eases the reconstruction, in order to facilitate feature representation learning in MAE; thereby, the method retains simplicity and efficiency of MAE [22]. In the experiment, the methods are analyzed from various aspects and are shown to improve performance on image-based MAE.

Table 6: Performance results on ADE-20K semantic segmentation by transferring the pre-trained ViT-B models.

Method	Epoch	mIOU	mAcc
MAE [22]†	800	42.60	54.07
Ours ($\sigma = 1$)	800	44.26	**55.85**
Ours ($\sigma = 4$)	800	**44.47**	55.76
MAE [22]†	1600	43.46	55.55
Ours ($\sigma = 1$)	1600	44.72	56.55
Ours ($\sigma = 4$)	1600	**45.33**	**57.08**

References

1. Bao, H., Dong, L., Piao, S., Wei, F.: Bert pre-training of image transformers. In: ICLR (2021)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: NeurIPS. pp. 1877–1901 (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
4. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML. pp. 1691–1703 (2020)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
6. Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., Wang, J.: Context autoencoder for self-supervised representation learning. IJCV **132**(1), 208–223 (2024)
7. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV. pp. 9640–9649 (2021)
8. Clark, P.J., Evans, F.C.: Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology **35**(4), 445–453 (1954)
9. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR. pp. 886–893 (2005)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186 (2019)
12. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., Guo, B.: Peco: Perceptual codebook for bert pre-training of vision transformers. In: AAAI. pp. 552–560 (2023)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Trans- formers for image recognition at scale. In: ICLR (2021)

14. Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., Brox, T.: Discriminative unsupervised feature learning with exemplar convolutional neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1734–1747 (2016)
15. Everett, R., Zupan, M.: Explorations into the mean nearest-neighbor distance in uniform and unimodal random distributions. Materials Today Communications **31**, 103637 (2022)
16. Everett, R., Zupan, M.: Mean nearest-neighbor distance in random 2d binary mixtures. Materials Today Communications **35**, 105882 (2023)
17. Feng, Z., Zhang, S.: Evolved part masking for self-supervised learning. In: CVPR. pp. 10386–10395 (2023)
18. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
19. Gonzalez, R., Woods, R.: Digital Image Processing. Prentice Hall (2007)
20. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent: A new approach to self-supervised learning. In: NeurIPS. pp. 21271–21284 (2020)
21. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR. pp. 1735–1742 (2006)
22. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
23. Hinojosa, C., Liu, S., Ghanem, B.: Colormae: Exploring data-independent masking strategies in masked autoencoders. In: ECCV. pp. 432–449 (2024)
24. Hondru, V., Croitoru, F.A., Minaee, S., Ionescu, R.T., Sebe, N.: Masked image modeling: A survey. IJCV **133**, 7154–7200 (2025)
25. Hu, R., Debnath, S., Xie, S., Chen, X.: Exploring long-sequence masked autoencoders (2022)
26. Jiang, L., Dai, B., Wu, W., Loy, C.C.: Focal frequency loss for image reconstruction and synthesis. In: ICCV. pp. 13919–13929 (2021)
27. Kong, L., Ma, M.Q., Chen, G., Xing, E.P., Chi, Y., Morency, L.P., Zhang, K.: Understanding masked autoencoders via hierarchical latent variable models. In: CVPR. pp. 7918–7928 (2023)
28. Krizhevsky, A., Hinton, G.E.: Learning multiple layers of features from tiny images. Technical report, University of Toronto (2009)
29. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: ICLR (2020)
30. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. In: NeurIPS. pp. 14290–14302 (2022)
31. Liu, H., Jiang, X., Li, X., Guo, A., Jiang, D., Ren, B.: The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. In: AAAI. pp. 1649–1656 (2023)
32. Liu, Y., et al: Pixmim: Rethinking pixel reconstruction in masked image modeling (2024)
33. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
34. Madan, N., Ristea, N.C., Nasrollahi, K., Moeslund, T.B., Ionescu, R.T.: What to hide from your students: Attention-guided masked image modeling. In: ECCV. pp. 300–318 (2022)
35. Madan, N., Ristea, N.C., Nasrollahi, K., Moeslund, T.B., Ionescu, R.T.: Cl-mae: Curriculum-learned masked autoencoders. In: WACV. pp. 2492–2502 (2024)
36. Nassar, J., Sokol, P., Chung, S., Harris, K.D., Park, I.M.: On 1/n neural representation and robustness. In: NeurIPS. pp. 6211–6222 (2020)

37. Nelsen, R.B.: An introduction to copulas. Springer, New York (2006)
38. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: ECCV. pp. 69–84 (2016)
39. Pathak, D., Girshick, R., Dollár, P., Darrell, T., Hariharan, B.: Learning features by watching objects move. In: CVPR. pp. 2701–2710 (2017)
40. Prates, M.O., Dey, D.K., Willig, M.R., Yan, J.: Transformed gaussian markov random fields and spatial modeling of species abundance. Spatial Statistics **14**(C), 382–399 (2015)
41. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831 (2021)
42. Roy, O., Vetterli, M.: The effective rank: A measure of effective dimensionality. In: EUSIPCO. pp. 606–610 (2007)
43. Shi, Y., Siddharth, N., Torr, P.H., Kosiorek, A.R.: Adversarial masking for self-supervised learning. In: ICML. pp. 20026–20040 (2022)
44. Torquato, S.: Mean nearest-neighbor distance in random packings of hard d-dimensional spheres. Physical Review Letters **74**(12), 2156–2159 (1995)
45. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: ICML. pp. 1096–1103 (2008)
46. Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., Zhang, Z.: Hard patches mining for masked image modeling. In: CVPR. pp. 10375–10385 (2023)
47. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: ICML. pp. 9929–9939 (2020)
48. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: CVPR. pp. 14668–14678 (2022)
49. Wu, Q., Ye, H., Gu, Y., Zhang, H., Wang, L., He, D.: Denoising masked autoencoders help robust classification. In: ICLR (2023)
50. Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y.S., Loy, C.C.: Masked frequency modeling for self-supervised visual pre-training. In: ICLR (2023)
51. 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: CVPR. pp. 9653–9663 (2022)
52. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. In: NeurIPS. pp. 27127–27139 (2022)
53. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV. pp. 649–666 (2016)
54. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., Zhang, L.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: CVPR. pp. 6881–6890 (2021)
55. Zheng, T., Li, B., Wu, S., Wan, B., Mu, G., Liu, S., Ding, S., Wang, J.: Mfae: Masked frequency autoencoders for domain generalization face anti-spoofing. IEEE Transactions on Information Forensics and Security **19**, 4058–4069 (2024)
56. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 5122–5130 (2017)