

ColorMAE: Exploring data-independent masking strategies in Masked AutoEncoders

Carlos Hinojosa , Shuming Liu , and Bernard Ghanem 

King Abdullah University of Science and Technology (KAUST)
<https://carloshinojosa.me/project/colormae>

Abstract. Masked AutoEncoders (MAE) have emerged as a robust self-supervised framework, offering remarkable performance across a wide range of downstream tasks. To increase the difficulty of the pretext task and learn richer visual representations, existing works have focused on replacing standard random masking with more sophisticated strategies, such as adversarial-guided and teacher-guided masking. However, these strategies depend on the input data thus commonly increasing the model complexity and requiring additional calculations to generate the mask patterns. This raises the question: *Can we enhance MAE performance beyond random masking without relying on input data or incurring additional computational costs?* In this work, we introduce a simple yet effective data-independent method, termed ColorMAE, which generates different binary mask patterns by filtering random noise. Drawing inspiration from color noise in image processing, we explore four types of filters to yield mask patterns with different spatial and semantic priors. ColorMAE requires no additional learnable parameters or computational overhead in the network, yet it significantly enhances the learned representations. We provide a comprehensive empirical evaluation, demonstrating our strategy’s superiority in downstream tasks compared to random masking. Notably, we report an improvement of 2.72 in mIoU in semantic segmentation tasks relative to baseline MAE implementations.

Keywords: Masked AutoEncoders · Data-independent masking · Masking strategy · Self-supervised learning · Masked Image Modeling

1 Introduction

Self-supervised learning (SSL) has emerged as a prominent pre-training paradigm, favored for its capacity to learn rich representations without the need for human-labeled data [19, 36]. Recent advancements demonstrate that large-scale SSL significantly outperforms supervised learning on challenging datasets. Inspired by masked language modeling (MLM) [5, 15] in natural language processing and the development of vision transformers (ViT) [17], masked image modeling (MIM) has achieved outstanding downstream performance across a broad spectrum of computer vision tasks [4, 24], thereby attracting increasing attention.

MIM learns rich representations during pre-training by masking certain patches of the input image and predicting their original content based on the remaining

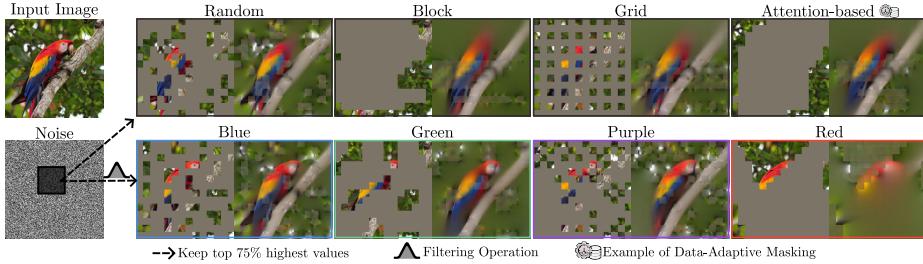


Fig. 1: We use the MAE [24] to mask and reconstruct an input image using different masking strategies with a masking ratio of 75%. The results of the first three columns shown in the top row correspond to traditional data-independent strategies: random, block-wise, and grid-wise masking, respectively. The last column of the top row shows an example of adaptive masking using an attention-based mechanism. The second row shows our four distinct types of masks generated by ColorMAE when filtering random noise with high-pass (Blue), band-pass (Green), band-stop (Purple), and low-pass (Red) filters.

unmasked patches. In this context, the strategy of masking plays a pivotal role. Current strategies fall into two categories: *data-independent masking* and *data-adaptive masking*. The former includes conventional random masking [24], block-wise masking [4], which masks contiguous blocks of image patches, and grid-wise masking, which keeps one out of every four patches, as shown in the first row of Fig. 1. To date, random masking has proven effective for most MIM methods due to its higher masking ratio and the implementation simplicity [24]. The latter category, *adaptive masking strategies*, involves designing mask patterns based on additional information, feedback, or image context analysis [27, 30], as illustrated in the first row, last column of Fig. 1. While adaptive masking often excels in downstream tasks, it necessitates the incorporation of attention-driven mechanisms or additional feedback, thereby increasing computational costs. Conversely, data-independent masking strategies, though simpler and without extra computational demands, have been scarcely explored beyond random masking.

In this work, we advocate for data-independent masking and propose a novel, straightforward, yet effective noise-filtering masking strategy. This strategy maintains the simplicity advantage of conventional random masking while facilitating the learning of stronger visual representations. Specifically, MAE initially generates a random noise array and selects the patches according to the desired mask ratio. From the signal/image processing perspective, the random noise array could be generated from white noise [7], characterized by a flat frequency spectrum with equal power across all bands. Considering noise frequency analysis and inspired by the concept of color noise in image processing [14, 29], we propose to generate mask patterns with varied constrained spectra, named ColorMAE. By filtering random noise through low-pass, high-pass, band-pass, and band-stop filters, we produce different noise patterns labeled as red, blue, green, and purple, respectively. These variants possess distinct properties and

frequency spectra, and show different spatial and semantic priors, as depicted in the second row of Fig. 1.

Our method, which does not rely on external guidance or additional learnable parameters, maintains computational efficiency during pre-training similar to random masking. Our extensive experiments reveal that one of our innovative masking strategies enables MAE to learn superior image representations during pre-training, surpassing conventional random masking in three downstream vision tasks: image classification, semantic segmentation, and object detection. Notably, our approach achieves a significant increase in mean Intersection over Union (mIoU) by 2.72 in semantic segmentation compared to random masking. We hope our approach could be used to design new MIM architectures or as a foundational masking strategy for developing adaptive masking techniques.

Contributions. We summarize our contributions as follows:

- (i) We propose a simple yet effective masking strategy to generate different data-independent masks by sampling and filtering random noise. Our method does not incorporate additional learnable parameters into the MAE model, preserving computational efficiency during pre-training.
- (ii) We investigate four distinct mask types created by applying low-pass, high-pass, band-pass, and band-stop filters to random noise. We offer detailed analysis and comparisons of these masks across three downstream tasks: image classification, semantic segmentation, and object detection.
- (iii) Through extensive experiments, we demonstrate that the “Green masking” ([ColorMAE-G](#)), achieved by applying a band-pass filter to random noise, significantly enhances MAE performance compared to random masking.

2 Related Works

2.1 Self-supervised Learning

Contrastive Learning. Among the numerous self-supervised learning approaches in computer vision that focus on learning from unlabeled data, contrastive learning has emerged in recent years [2, 12, 23, 26, 36, 45]. Its fundamental principle involves learning representations through instance discrimination, which involves attracting similar samples while optionally repelling dissimilar ones. SimCLR [10], a prominent method in this domain, enhances representations by maximizing the similarity between different views of the same image with large training batches. MoCo [25] advances this approach by employing a memory bank and a momentum-updated encoder to complement the pool of negative samples, thereby learning more robust representations. DINO [6] incorporates a self-distillation mechanism, compelling the student network to mimic the teacher network’s output on augmented views of the same image, which fosters strong attention to the salient parts of images.

Masked Image Modeling. Inspired by the success of Masked Language Modeling (MLM) [15] in natural language processing, Masked Image Modeling (MIM)

has garnered increasing interest in the vision domain [3, 11, 18, 24, 33, 51]. This approach aims to reconstruct the original image from masked inputs, with reconstruction targets varying from raw pixels [8, 24, 48] and dVAE tokens [4] to HoG features [43] and frequency components [47]. Notably, MAE [24] has achieved significant attention for its simplicity and computational efficiency. It introduces an asymmetric encoder-decoder architecture, where the encoder processes only a subset of visible patches selected through a random masking strategy, while a lightweight decoder predicts the original image using both masked and unmasked patches. Pretraining with MAE has been shown to substantially enhance performance on downstream tasks compared to supervised pretraining. This work builds upon MAE’s framework and proposes novel masking strategies that improve the random masking approach.

2.2 Masking Strategy

Data-Adaptive Masking. The selection of mask sampling strategies is pivotal in MIM as it defines the difficulty of the pretext task, thereby influencing both the quality of the reconstruction and the learned representations. Recent studies have explored more advanced masking strategies. For example, AttMask [27] selects patches for masking based on high scores in the teacher network’s attention map, presenting a more challenging pretext task. ADIOS [38] utilizes adversarial training to increase the difficulty of the pretext task. SemMAE [30] targets semantic portions of the image, masking patches within these areas to provide a nuanced challenge. HPM [42] posits that the difficulty of MIM reconstruction can be quantified by patch-wise reconstruction loss, leading to the development of an auxiliary loss predictor for strategic masking. Feng [21] introduces an evolved masking strategy that incrementally focuses on object semantics and context by effectively masking precise object parts. These strategies, which depend on the image’s pixel values for mask sampling, are collectively referred to as *data-adaptive masking*. This term signifies that the masking process is conditioned and adaptively chosen based on the input data.

Data-Independent Masking. Conversely, a category of masking techniques exists that does not depend on the input images or external guidance, termed *data-independent masking*. Among these, random masking stands out for its simplicity and is implemented in MAE [24] and SimMIM [48] with a large masking ratio. Block-wise masking, adopted by BEiT [4] and BootMAE [16], involves masking contiguous blocks of image patches. Furthermore, grid masking—a strategy that regularly obscures a grid pattern across the image—has been proposed as a data augmentation method in [9] and explored in MAE. Our work extends the data-independent masking framework by applying different filters on random noise, thus achieving computational efficiency with a fast-speed masking function. This proposed methodology not only enhances visual representation learning but also provides substantial benefits for downstream tasks, offering an improvement over traditional random masking approaches.

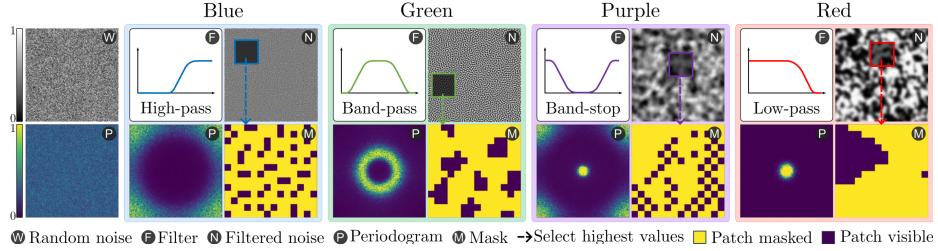


Fig. 2: Starting with random noise \mathbb{W} , we apply four filters \mathbb{F} : high-pass (Blue), band-pass (Green), band-stop (Purple), and low-pass (Red) to produce the filtered noises \mathbb{N} . The periodogram \mathbb{P} of each filtered noise is displayed in the second row; as observed, each version of \mathbb{N} exhibits a distinctive pattern in the frequency domain. Once the filtered noise is obtained, we perform a random crop on \mathbb{N} to obtain a local window (sized to match the total number of patches) and select the top values according to the desired mask ratio (e.g., 75%) to create the binary mask \mathbb{M} used during pre-training.

3 Proposed Method

To efficiently create binary masks during pre-training, MAE [24] generates random uniform noise and then selects the data points with the highest top values according to the desired mask ratio (e.g., 75%). These selected values determine which patches are masked out (represented by ones in the binary mask) or remain visible in the input image, hence directing which portions of the input data the model will attempt to reconstruct. In this regard, the mask is critical in determining how high-level semantic representations are extracted from low-level features like image pixels.

In addition to uniform noise, *white noise* [7] can also be implemented to produce random patterns. In general, uniform and white noise are not identical, as their mathematical definitions and statistical properties differ. For instance, white noise has an evenly distributed signal across the frequency spectrum, while uniform noise does not inherently have a flat spectral distribution. However, in practical experiments, they exhibit similar behavior and generate comparable random masks. Consequently, there are no significant differences in MAE pre-training and fine-tuning performance when incorporating either noise type; see our supplementary material for further details. Therefore, we will refer to both white and uniform noise simply as *random noise* throughout the manuscript.

Moreover, beyond white noise, there are other types of noise in the image processing field, known as *color noise* [14, 29, 40]. Unlike white noise, which maintains consistent power across its frequency bands, color noises exhibit unique spectral distributions, such as a predominance in the low-frequency band. Although color noise has been investigated in other domains, its application within deep learning frameworks, especially self-supervised learning, remains largely unexplored.

In this work, we draw inspiration from the concept of color noise in image processing and introduce a novel approach, termed **ColorMAE**, which employs mask patterns with distinct spectral constraints to facilitate efficient self-supervised

learning. Concretely, instead of directly using random noise, our method applies different filters to random noise, *e.g.* low-pass, high-pass, band-pass, and band-stop, to generate diverse noise patterns that embody unique spatial and frequency characteristics, as illustrated in Fig. 2. This masking strategy is independent of data, eliminating the need for additional instructional inputs or loss functions, thus ensuring efficient and rapid sampling akin to MAE’s random masking. However, our experiments demonstrate that certain color noise can significantly enhance visual representation quality during pre-training. To align with traditional terminology in image processing, we categorize the produced noise patterns as **Red**, **Blue**, **Green**, and **Purple** noise. Next, we will formally detail the definitions and implementations of these color noise masking.

Red Noise. Let $W(x, y)$ represent a random noise image, where x and y are spatial coordinates. We apply a blurring operation over W using a Gaussian kernel G_σ with standard deviation σ to filter out the high-frequency components and accentuate low frequencies effectively. This operation transforms the random noise into *red noise* N_r given by:

$$N_r = G_\sigma * W, \quad (1)$$

where $*$ denotes the convolution operation. Finally, we perform normalization on N_r to adjust the intensity values accordingly. We iteratively repeat the low-pass filtering and normalization steps to refine the noise characteristics.

Blue Noise. To generate blue noise patterns, it is required to apply a high-pass filter over W . A practical approach to implementing a high-pass filter involves first applying a low-pass filter ($G_\sigma * W$) to obtain the low-frequency content. Then, this filtered output is subtracted from the original random noise image W , effectively retaining the high-frequency components. The resulting *blue noise* N_b is formally expressed as

$$N_b = W - G_\sigma * W. \quad (2)$$

Note that alternative algorithms, such as the Void and Cluster method [39], can also be employed to generate high-quality blue noise patterns. This algorithm initiates with a random distribution of points and iteratively adjusts their placement to fill gaps evenly while avoiding the formation of clusters. It computes a density metric for empty spaces based on nearby points, placing new points strategically for even distribution. This algorithm and the blue noise patterns have been extensively used in the computer graphics field [1, 44].

Green Noise. This noise is defined as the mid-frequency component of white noise; *i.e.*, it can be generated by applying a band-pass filter over W to eliminate both high and low frequencies. Such band-pass filtering effect can be approximated by sequentially applying two Gaussian blurs: first, a weak blur is applied to W to remove the highest frequency details, followed by a separate strong blur to capture the lowest frequency content of W . By subtracting the strongly blurred version of W from the weakly blurred one, the resultant noise image

retains only the mid-frequency components. Formally, the *green noise* N_g image can be obtained as:

$$N_g = G_{\sigma_1} * W - G_{\sigma_2} * W, \quad (3)$$

where σ_1 and σ_2 denote the standard deviation of the two Gaussian kernels with $\sigma_1 < \sigma_2$.

Purple Noise. Finally, in this paper, we refer to purple noise as the noise that has only high and low-frequency content, *i.e.*, does not have a middle-frequency component. We apply a band-stop filter over the random noise W to produce this type of noise. Specifically, we first apply a band-pass filter to W to obtain green noise and then subtract it from the input W , preserving only the low and high frequencies. Formally, this transformation of the noise W into *purple noise* N_p can be expressed as:

$$N_p = W - (G_{\sigma_1} * W - G_{\sigma_2} * W), \quad (4)$$

where $\sigma_1 < \sigma_2$. Analyzing the periodogram in Fig. 2 (column “Purple”), we can observe that this noise combines the characteristics of both red and blue noise.

Mask Generation. In implementation, we pre-compute *color noise* offline and store them in GPU memory before initiating MAE pre-training. To efficiently generate the masks during pre-training, we first apply random transformations on the loaded noise tensor to get a P -sized square noise window for every image in the batch B , where P is the total number of patches. Then, we select the highest values from the noise window according to the desired mask ratio. Specifically, we apply random crop, horizontal flip, and vertical flip image transformation. Note that these image transformations operate in the spatial domain; hence, the frequency properties described in the previous section are preserved [22]. Algorithm 1 shows the pseudo-code for our masking approach in PyTorch style.

Figure 2 shows examples of the generated masks for each color noise (see ⑩). As observed, the produced masks have a particular pattern associated with the frequency properties of each noise. For example, blue noise produces binary masks whose values are distributed uniformly but without large empty areas or overly dense clusters. On the other hand, the masks produced by green noise can be seen as a “clustered” version of the masks produced with blue noise.

4 Experiments

Implementation Details. We evaluate the performance of our proposed masking strategies under self-supervised pre-training with MAE [24] on the ImageNet-1K [37] dataset. Unless otherwise specified, we mainly use the standard ViT-B/16 [17] as the backbone, and the decoder consists of 8 Transformer layers with a hidden dimension 512. The input images are resized to 224×224 , and the patch size is 16×16 ; thus, the resulting total sequence length is $L = 196$. To make a fair comparison with the original MAE pre-trained with random

Algorithm 1: Pseudo-Code of our masking approach in PyTorch style.

```

import torch

def mask_generation(N, P, B, T, mask_ratio):
    # N: noise tensor (e.g. blue, green, purple or red noise).
    # P: total number of patches. B: batch size.
    # T: random crop, horizontal, and vertical flip PyTorch transforms.
    # mask_ratio: the mask ratio of total patches (e.g. 0.75).
    # apply random transforms (T) to get a  $\sqrt{P} \times \sqrt{P}$  noise windows
    windows = T(N)[:, :B] # Assuming B < N.shape[0]
    len_keep = int(P * (1 - mask_ratio))
    windows = windows.view(B, -1)

    # keep stronger values from the noise
    ids_shuffle = torch.argsort(windows, dim=1, descending=True)
    ids_restore = torch.argsort(ids_shuffle, dim=1)
    ids_keep = ids_shuffle[:, :len_keep]

    # generate the binary mask: 0 is keep, 1 is remove
    mask = torch.ones([B, P])
    mask[:, :len_keep] = 0

    # unshuffle to get the binary mask
    mask = torch.gather(mask, dim=1, index=ids_restore)
    return mask, ids_restore, ids_keep

```

masking [24], we use the same masking ratio of 75%. Please refer to our supplementary material for additional details on implementation and experiments with different masking ratios. We evaluate transfer learning performance using our pre-trained **ColorMAE** models on different datasets and downstream tasks described as follows:

ImageNet Classification. We evaluate the performance of MAE pre-trained with our proposed masking strategy on ImageNet-1K [37] classification following the standard protocol [24]. We perform end-to-end fine-tuning for 100 epochs and report the Top-1 accuracy (%) obtained on the validation set. We maintain the same resolution of 224×224 on both pre-training and fine-tuning.

COCO Object Detection and Instance Segmentation. We employ ViT-Det [32] as our object detector model, which utilizes a Vision Transformer backbone to perform object detection and instance segmentation. Unless otherwise specified, we perform end-to-end fine-tuning on the COCO dataset [34], resizing the images to a resolution of 768×768 to expedite the fine-tuning process. We report the box average precision (AP^{bbox}) for object detection and the mask AP for instance segmentation (AP^{mask}).

ADE20k Semantic Segmentation. We employ UperNet [46] as our segmentation model and perform end-to-end fine-tuning on the ADE20k [50] dataset for 160k iterations with an image resolution of 512×512 . The evaluation metric used is the mean Intersection over Union (mIoU) [20].

4.1 Exploring Masking Strategies Performance

Qualitative Results. In this section, we evaluate the performance of MAE on the downstream tasks mentioned in the previous section when we pre-train with our proposed four types of **ColorMAE** masks. Figure 3 presents visualizations

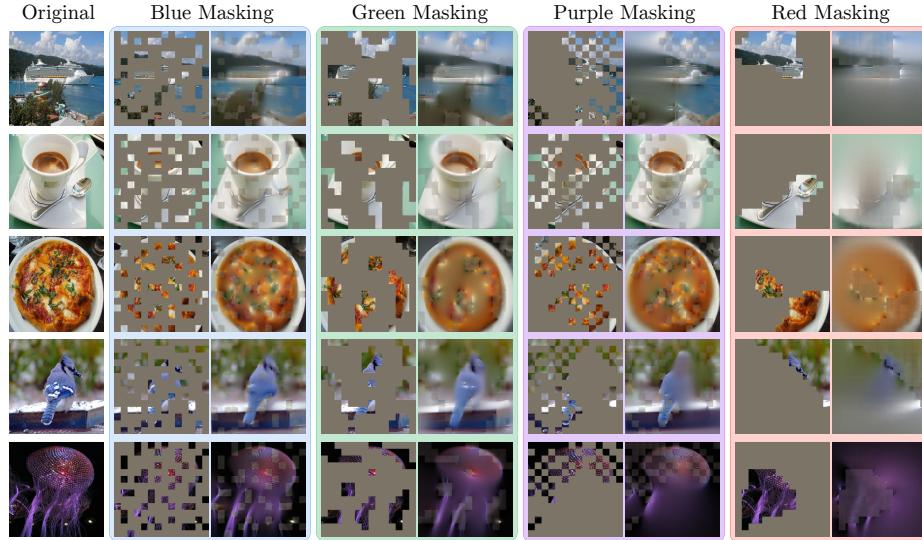


Fig. 3: Reconstruction results on ImageNet validation images from MAE pre-trained during 300 epochs with our four generated masks: Blue, Green, Purple, and Red.

Table 1: Downstream tasks performance after fine-tuning. MAE is pre-trained on ImageNet-1K [37] with random masking and our proposed masking approach. We report ImageNet-1K Top-1 accuracy, ADE20K mIoU [50], and COCO AP^{bbox} [34] for classification, semantic segmentation, and object detection, respectively.

Pretrain Epochs	Classification (Top-1 accuracy)					Semantic Segmentation (mIoU)					Object Detection (AP ^{bbox})				
	Random	Blue	Green	Purple	Red	Random	Blue	Green	Purple	Red	Random	Blue	Green	Purple	Red
100	81.69	81.82	81.82	80.82	78.83	42.20	40.33	42.24	38.22	35.31	45.90	46.00	45.90	44.10	40.80
300	82.82	82.56	82.98	82.39	81.35	44.51	43.42	45.80	43.85	42.08	48.50	48.10	48.70	47.20	45.10
800	83.17	83.02	83.57	82.92	82.41	46.46	44.81	49.18	45.96	44.78	49.15	49.10	49.50	48.50	46.90
1600	83.43	83.26	83.77	83.20	82.73	47.46	46.35	49.26	47.23	46.08	49.60	49.50	50.10	49.10	47.20

of the ImageNet validation images reconstructed with our four types of masks: Blue, Green, Purple, and Red. All the masks have the same masking ratio of 75%, and the MAE model is pre-trained for 300 epochs to reconstruct unnormalized pixels. As observed, Blue masking provides better reconstruction quality as the reconstruction task is easy for the decoder. Similarly, while Green masking yields a lower reconstruction quality than blue, it still provides semantically meaningful and sufficiently accurate reconstructions. Conversely, the reconstruction task becomes significantly more challenging with Purple and Red masking, resulting in lower reconstruction quality. Specifically, Red masking is notably more “aggressive”, leading to poor reconstruction and representation learning. Please refer to our supplementary for more reconstruction visualizations.

Quantitative Results. In Tab. 1, we investigate the performance of our four types of masks generated by our approach on various downstream tasks and show the comparison with traditional random masking. Our findings indicate

Table 2: Additional MAE experiments with (a) ViT Large (ViT-L/16 [17]) as backbone and evaluated on two downstream tasks: ImageNet-1K classification and ADE20K semantic segmentation; (b) ViT Base (ViT-B/16 [17]) as backbone and evaluated on COCO object detection when using images with 768×768 and 1024×1024 resolution.

Pretrain Epochs	Arch	ImageNet-1K (Top 1 Acc)			ADE20K (mIoU)			Image Size 768x768			Image Size 1024x1024		
		Random	Blue	Green	Random	Blue	Green	Random	Blue	Green	Random	Blue	Green
300	ViT Large	84.76	84.77	85.02	47.55	46.75	49.00	50.29	49.38	51.46	48.50	48.10	48.70
800	Large	85.42	85.34	85.64							50.10	49.80	50.40

(a) Downstream tasks performance with MAE pre-trained using ViT-Large as the backbone and using random, Blue, and Green masking.

(b) Object detection performance when fine-tuning using images resized to 768×768 vs. images with 1024×1024 resolution.

that the Purple and Red masks exhibit lower performance, with the latter being the worst, unable to learn representations effectively. This aligns with the visualizations in Fig. 3, where it clearly shows that Red masking significantly increases the difficulty of the reconstruction task. Note that although Blue masking provides the best (lower) reconstruction loss among all the approaches (Fig. 4), it does not yield better performance on downstream tasks (Tab. 1). Finally, it is important to highlight that Green masking delivers the best results across all the evaluated downstream tasks. The improvement is particularly significant in the semantic segmentation task, with a notable increase of 2.72 in the mIoU metric compared to random masking when pre-training MAE for 800 epochs. We also observe that while Green masking consistently provides performance improvements, such enhancements become more evident when the pre-training epochs increase. Notably, there is a marked improvement in mIoU ($3.38\uparrow$) when increasing the training from 300 to 800 epochs, which suggests our approach provides faster convergence than traditional random masking.

Additional Results. Tab. 2 (a) presents additional experiments when pre-training MAE using ViT Large (ViT-L/16 [17]) as a backbone and evaluate downstream performance on ImageNet-1K classification task and semantic segmentation on ADE20K dataset. The table shows that the results are consistent with those presented in Tab. 1. Specifically, Green masking performs best in both tasks, outperforming random and Blue masking. Additionally, Tab. 2 (b) presents a comparison between object detection performance when fine-tuning on the COCO dataset using images resized to 768×768 resolution versus 1024×1024 resolution. This experiment uses ViT Base (ViT-B/16 [17]) as the backbone for MAE. The performance enhancements obtained with Green masking are consistent with previous results, outperforming random and Blue masking.

4.2 Comparison with Other Methods

Given the results from the previous section, we will focus this section on comparing Green masks with other methods. Here, we refer to MAE pre-trained with Green masks as [ColorMAE-G](#).

Comparison with MAE. The bottom block of Tab. 3 showcases that our [ColorMAE-G](#) consistently outperforms MAE[†] (our implementation), as well as the results presented in the original MAE paper [24], without incurring additional

Table 3: Comparison with state-of-the-art methods pre-trained on ImageNet-1K. The resolution of images is 224×224 for both pre-training and fine-tuning. \dagger indicates our implementation, including pre-training and fine-tuning. \ddagger means the results are borrowed from [11]. \S means the results are borrowed from [41].

Method	Pretrain Epoch	Pre-trained Data	ADE20K		ImageNet		COCO					
			mIoU	Top-1 Acc.	AP ^{bbox}	AP ₅₀ ^{bbox}	AP ₇₅ ^{bbox}	AP ^{mask}	AP ₅₀ ^{mask}	AP ₇₅ ^{mask}		
<i>Non-MIM</i>												
MoCo v3 [‡] [13]	600	IN1K	47.2	83.0	45.5	67.1	49.4	40.5	63.7	43.4		
DINO [‡] [6]	1600	IN1K	47.2	83.3	46.8	68.6	50.9	41.5	65.3	44.5		
DropPos [41]	800	IN1K	47.8	84.2	47.7	68.3	52.8	42.6	65.3	46.2		
<i>MIM with data-adaptive masking</i>												
AttMask [27]	100	IN1K	45.3	-	48.8	-	-	42.0	-	-		
UM-MAE [31]	200	IN1K	42.6	82.9	45.9	64.5	50.2	-	-	-		
SemMAE [§] [30]	800	IN1K	44.9	83.4	45.6	66.2	55.2	40.9	63.3	44.4		
HPM [42]	800	IN1K	48.5	84.2	50.1	-	-	44.6	-	-		
<i>MIM with data-independent masking</i>												
BEiT [4]	800	IN1K+DALLE	45.6	83.2	40.8	59.4	44.1	36.0	56.8	38.2		
MAE [†] [24]	800	IN1K	46.5	83.2	49.2	69.7	53.9	43.4	66.6	46.9		
MixedAE [8]	800	IN1K	48.7	83.5	50.3	69.1	54.8	43.5	66.2	47.4		
ColorMAE-G	800	IN1K	49.2	83.6	49.5	70.0	54.2	43.7	67.1	47.1		
MAE [24]	1600	IN1K	48.1	83.6	50.6	69.4	55.0	43.8	66.6	47.5		
ColorMAE-G	1600	IN1K	49.3	83.8	50.1	70.7	54.7	44.4	67.8	48.0		

overhead (see Tab. 4). The most notable enhancement is observed in the semantic segmentation task, where there is a significant improvement of $+2.7$ mIoU with 800 pretraining epochs. Furthermore, our method also demonstrates competitive or superior results in object detection tasks.

Comparison with Other Data-Independent Masking Methods. In the last block of Tab. 3, we also compare our approach against other state-of-the-art data-independent masking methods, which usually use random or block-wise masking. As observed, ColorMAE-G outperforms all these methods in the semantic segmentation task on the ADE20K dataset and provides comparable results in the other downstream tasks, where it is only surpassed by MixedAE (-0.2) and MAE (-0.5) in COCO object detection.

Comparison with Data-adaptive Masking Methods. Notably, our data-independent masking approach also achieves competitive performance even compared to sophisticated data-adaptive masking, which incorporates additional attention-based or adversarial-guided mechanisms, increasing the computational cost. In the semantic segmentation task on the ADE20K dataset, our approach visibly outperforms these methods. Our performance in the object detection task is also better than these approaches and comparable to HPM [42]. However, we do not introduce any additional parameters or computations in the network, thus enjoying the benefit of fast training (see Tab. 4).

4.3 Analysis

Reconstruction Loss vs. Downstream Performance. Figure 4 presents the MAE pre-training (reconstruction) loss curves for random masking and our

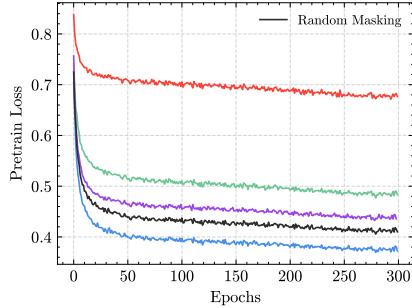


Fig. 4: MAE pre-training loss for different masking strategies with ViT-B.

Table 4: Complexity analysis of the MAE model when pre-training with traditional random masking and our proposed masking strategies.

Masking Strategy	Parameters (M)	Flops (G)	Memory (GB)	Pre-training Time per Epoch (Min)
Random	111.91	16.87	27.44	5.21
Blue	111.91	16.87	28.21	5.18
Green	111.91	16.87	28.21	5.18
Purple	111.91	16.87	28.21	5.18
Red	111.91	16.87	28.21	5.18

proposed masking approach. As observed, Blue masking achieves the lowest reconstruction loss over the epochs, followed by random, Purple, Green, and Red masking. Interestingly, Green masking does not yield the lowest pre-training loss, yet it achieves the best performance in downstream tasks, see Tab. 1. Our results contradict the hypothesis that a lower reconstruction loss implies better downstream performance [49]. However, our findings resonate with the observations in [42]. Specifically, considering the reconstruction loss as a metric of the difficulty of the pre-training task, authors in [42] propose to mask the patches with higher loss, increasing downstream performance. Such hard-to-reconstruct patches are usually associated with the discriminative parts of an image, like objects. Authors in [42] conclude that consistently increasing the difficulty of the pretext task does not lead to better performance, and retaining a certain degree of *randomness* is necessary for better results. Similarly, as observed from Fig. 4, our Red masking approach tends to mask out big segments of the image, making the pretext task very difficult but not allowing the model to learn useful feature representations. On the other hand, our Green masking approach masks out smaller random segments in the image, making the pretext task difficult enough to learn better representations. In general, our Green masking provides a better balance between pre-training task difficulty and randomness.

Computational Cost. In general, data-adaptive masking approaches inevitably increase the computational cost and number of parameters since they need to introduce additional components to the network, *e.g.*, an extra decoder. For instance, HPM [42] increases the training time $1.1\times$, while CAE [11] increases the number of parameters to $1.23\times$ and training time to $1.24\times$ in comparison with MAE [24]. Similarly, authors in [21] report that their mask generation occupies 12% of pre-training time. On the other hand, our proposed data-independent masking strategy is efficient and does not add extra model parameters or computational overhead, as shown in Tab. 4. Because we pre-compute the noise color patterns offline and store them in GPU memory, there is only a small increment in memory usage compared to the original MAE model. In particular, during our experiments, we use 3072 noise patterns of 256×256 spatial dimension for each

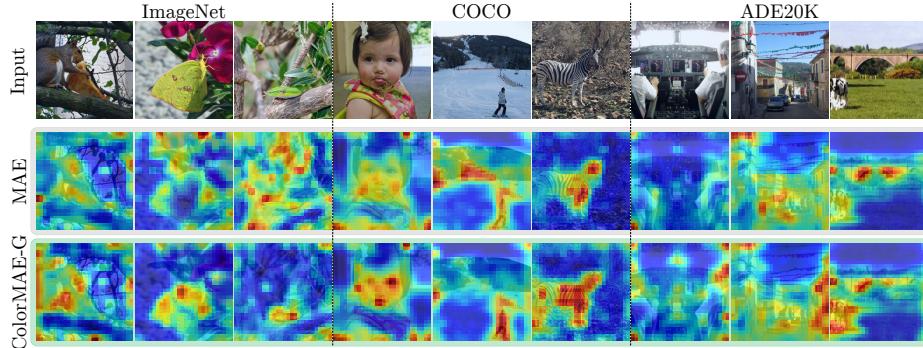


Fig. 5: Self-attention of the [CLS] tokens averaged across the heads of the last layer in MAE pre-trained using random masking and our proposed Green masking approach (**ColorMAE-G**). We show attention maps on images from Imagenet-1K [37](1st-3rd columns), Microsoft COCO [34](4th-6th columns) and ADE20K [50](7th-9th columns) datasets. Both MAE and **ColorMAE-G** are pre-trained on ImageNet-1K for 300 epochs. Please refer to our supplementary for more visualizations of the attention maps when pre-training MAE with other **ColorMAE** masks.

type of color noise, leading to a small increment of 2.8% of memory. However, fewer patterns can be used, reducing memory costs while slightly impacting performance. Please see our supplementary for additional experiments when varying the number of noise patterns used during **ColorMAE** pre-training.

Attention Analysis. We show examples of self-attention maps of the [CLS] tokens averaged across the heads of the last layer in Fig. 5 for the three different datasets. Here, we show the results for MAE pre-trained using random and our proposed Green masking approach. From the visualizations, our **ColorMAE-G** effectively identifies the foreground object patches with better precision and completeness. This might also explain its superior performance when transferred to dense perception tasks such as semantic segmentation [50], object detection, and instance segmentation [34].

Additionally, in Fig. 6, we employed EigenCAM [35] to visualize the activation patterns of ViT-B, highlighting the model focus areas during image classification tasks. We first perform self-supervised pre-training on the ImageNet-1K dataset using MAE with random masking and our **ColorMAE-G** approach. Then, we conduct end-to-end supervised fine-tuning for 100 epochs on the ImageNet classification task. The resulting CAMs, depicted in the second and third rows of Fig. 6, offer a visual comparison of the attention mechanisms of the models. As observed, the CAMs of ViT-B pre-trained with our **ColorMAE-G** exhibit more localized and relevant attention areas (*e.g.*, discriminative objects/subjects in the scene), especially in contrast to those provided by ViT pre-trained with the standard MAE. Please refer to our supplementary document for more self-attention and CAM maps.

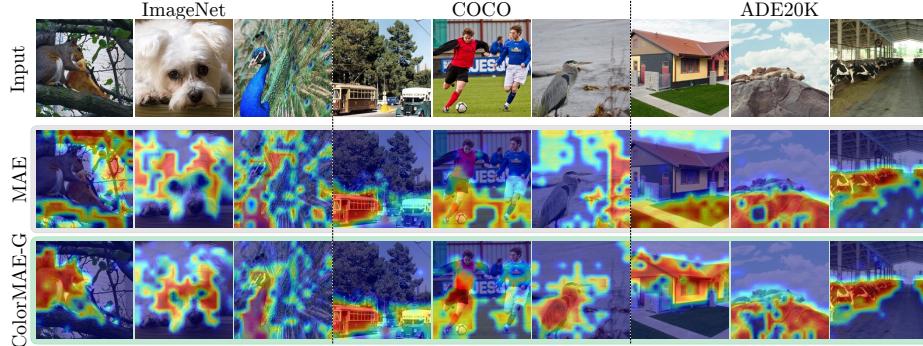


Fig. 6: Comparative visualization of Class Activation Maps (CAM) generated with EigenCAM [35] for ViT-B. First, we perform self-supervised pre-training using standard MAE with random masking and our **ColorMAE-G** on the ImageNet-1K dataset. Then, we conduct end-to-end supervised fine-tuning following the standard protocol [24] for ImageNet classification during 100 epochs. We show CAM maps of the ViT-B pre-trained with MAE (second row) and our **ColorMAE-G** (third row) on images from ImageNet-1K (1st-3rd column), Microsoft COCO (4th-6th columns), and ADE20K (7th-9th columns) datasets. Please refer to our supplementary for more visualizations.

5 Conclusions

Until now, random masking has been the foundational strategy and a common starting point for developing data-adaptive masking strategies. This paper explored four distinct data-independent masking alternatives to the conventional random masking approach. Using our **ColorMAE** approach, we can generate different random masks with specific patterns by using noise with different frequency spectra. We observed that by using our generated *Green masks* during MAE pre-training (**ColorMAE-G**) we achieved faster convergence and better performance in downstream tasks, especially in semantic segmentation. Among the explored data-independent approaches in this paper, we found that our Green masks provide the best balance between pretext task difficulty and randomness, which allows the model to learn better representations.

Discussion and Limitations. In this work, we adopted a simple yet effective approach to filter random noise and generate different masks. While we investigated additional algorithms, such as the void-and-cluster method [39], for generating improved blue noise patterns, these did not yield significant performance enhancements and resulted in slower mask generation. Similarly, developing or using other algorithms to produce better *green noise* patterns [28] can be explored in future works. On the other hand, while our method is computationally efficient, it increments memory usage by 2.8%. Although this increment is small, it could be considered a limitation and could be improved in future works.

Acknowledgments. This work was supported by the KAUST Center of Excellence on GenAI under award number 5940.

References

1. Ahmed, A.G., Wonka, P.: Screen-space blue-noise diffusion of monte carlo sampling error via hierarchical ordering of pixels. *ACM Transactions on Graphics (TOG)* **39**(6), 1–15 (2020)
2. Bachman, P., Hjelm, R.D., Buchwalter, W.: Learning representations by maximizing mutual information across views. *Advances in neural information processing systems* **32** (2019)
3. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: *International Conference on Machine Learning*. pp. 1298–1312. PMLR (2022)
4. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9650–9660 (2021)
7. Castleman, K.R.: *Digital image processing*. Prentice Hall Press (1996)
8. Chen, K., Liu, Z., Hong, L., Xu, H., Li, Z., Yeung, D.Y.: Mixed autoencoder for self-supervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 22742–22751 (2023)
9. Chen, P., Liu, S., Zhao, H., Jia, J.: Gridmask data augmentation. arXiv preprint arXiv:2001.04086 (2020)
10. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: *International conference on machine learning*. pp. 1597–1607. PMLR (2020)
11. 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. *International Journal of Computer Vision* **132**(1), 208–223 (2024)
12. Chen, X., He, K.: Exploring simple siamese representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 15750–15758 (2021)
13. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9640–9649 (October 2021)
14. Correa, C.V., Arguello, H., Arce, G.R.: Spatiotemporal blue noise coded aperture design for multi-shot compressive spectral imaging. *JOSA A* **33**(12), 2312–2322 (2016)
15. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
16. Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: Bootstrapped masked autoencoders for vision bert pretraining. In: *European Conference on Computer Vision*. pp. 247–264. Springer (2022)
17. 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: Transformers for image recognition at scale. In: *9th International Conference on Learning Representations, ICLR 2021, Virtual*

- Event, Austria, May 3-7, 2021. OpenReview.net (2021), <https://openreview.net/forum?id=YicbFdNTTy>
18. El-Nouby, A., Izacard, G., Touvron, H., Laptev, I., Jegou, H., Grave, E.: Are large-scale datasets necessary for self-supervised pre-training? arXiv preprint arXiv:2112.10740 (2021)
 19. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: International Conference on Machine Learning. pp. 3015–3024. PMLR (2021)
 20. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015)
 21. Feng, Z., Zhang, S.: Evolved part masking for self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10386–10395 (2023)
 22. Gonzalez, R.C., Woods, R.E.: Digital image processing. Prentice Hall (2008)
 23. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Dondersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
 24. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
 25. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
 26. Hjelm, R.D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., Bengio, Y.: Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670 (2018)
 27. Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., Komodakis, N.: What to hide from your students: Attention-guided masked image modeling. In: European Conference on Computer Vision. pp. 300–318. Springer (2022)
 28. Lau, D.L., Arce, G.R., Gallagher, N.C.: Green-noise digital halftoning. Proceedings of the IEEE **86**(12), 2424–2444 (1998)
 29. Lau, D.L., Ulichney, R., Arce, G.R.: Blue and green noise halftoning models. IEEE Signal Processing Magazine **20**(4), 28–38 (2003)
 30. Li, G., Zheng, H., Liu, D., Wang, C., Su, B., Zheng, C.: Semmae: Semantic-guided masking for learning masked autoencoders. Advances in Neural Information Processing Systems **35**, 14290–14302 (2022)
 31. Li, X., Wang, W., Yang, L., Yang, J.: Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. arXiv preprint arXiv:2205.10063 (2022)
 32. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision. pp. 280–296. Springer (2022)
 33. Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., et al.: Mst: Masked self-supervised transformer for visual representation. Advances in Neural Information Processing Systems **34**, 13165–13176 (2021)
 34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–

- ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
35. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: 2020 international joint conference on neural networks (IJCNN). pp. 1–7. IEEE (2020)
 36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
 37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)
 38. Shi, Y., Siddharth, N., Torr, P., Kosiorek, A.R.: Adversarial masking for self-supervised learning. In: International Conference on Machine Learning. pp. 20026–20040. PMLR (2022)
 39. Ulichney, R.A.: Void-and-cluster method for dither array generation. In: Human Vision, Visual Processing, and Digital Display IV. vol. 1913, pp. 332–343. SPIE (1993)
 40. Vasseur, D.A., Yodzis, P.: The color of environmental noise. Ecology **85**(4), 1146–1152 (2004)
 41. Wang, H., Fan, J., Wang, Y., Song, K., Wang, T., ZHANG, Z.X.: Droppos: Pre-training vision transformers by reconstructing dropped positions. Advances in Neural Information Processing Systems **36** (2024)
 42. Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., Zhang, Z.: Hard patches mining for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10375–10385 (2023)
 43. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self-supervised visual pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14668–14678 (2022)
 44. Wolfe, A., Morrical, N., Akenine-Möller, T., Ramamoorthi, R., Ghosh, A., Wei, L.: Spatiotemporal blue noise masks. In: Eurographics Symposium on Rendering. pp. 117–126 (2022)
 45. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
 46. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
 47. Xie, J., Li, W., Zhan, X., Liu, Z., Ong, Y.S., Loy, C.C.: Masked frequency modeling for self-supervised visual pre-training. In: The Eleventh International Conference on Learning Representations (2022)
 48. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9653–9663 (2022)
 49. Zhang, Q., Wang, Y., Wang, Y.: How mask matters: Towards theoretical understandings of masked autoencoders. Advances in Neural Information Processing Systems **35**, 27127–27139 (2022)
 50. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
 51. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021)