Supplementary Material
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

Introduction

This supplementary document provides additional experiments, visualizations, and implementation details of our work. Specifically, we include the following:

1. Implementation Details
2. Varying Masking Ratio
3. Varying Number of Patterns
4. White Noise *vs*. Uniform Noise
5. Reconstruction Visualizations
6. Self-attention Maps Visualizations
7. CAM Visualizations

1 Implementation Details

We generate the noise patterns offline by applying low-pass, high-pass, band-pass, and band-stop filters to random noise. These filtering operations were implemented in C/C++ code. The noise patterns are created as 2D images, which are then concatenated into a NumPy array and will utilized during the MAE pre-training process. For green and purple noise, we selected the standard deviations of the Gaussian kernel as $\sigma_1 = 0.5$ and $\sigma_2 = 2.0$, respectively. For blue and red noise, σ was randomly chosen for each 2D generated image between 0.5 and 2.0. These values were empirically found to be optimal in our experiments, as variations in σ did not yield significant differences in performance. Future work may further investigate the selection methodology for these parameters and explore enhanced methods for generating color noise patterns.

During pre-training, masks were generated using Algorithm 1 (in the main paper) with PyTorch. All experiments were conducted on 8 Nvidia A100-80G GPUs for pre-training and fine-tuning, except for the semantic segmentation task, which utilized 4 GPUs. To ensure a fair comparison, we used the same 75% masking ratio as the original MAE pre-trained [1]. Please refer to the next sections for results when varying the masking ratio and number of noise patterns.

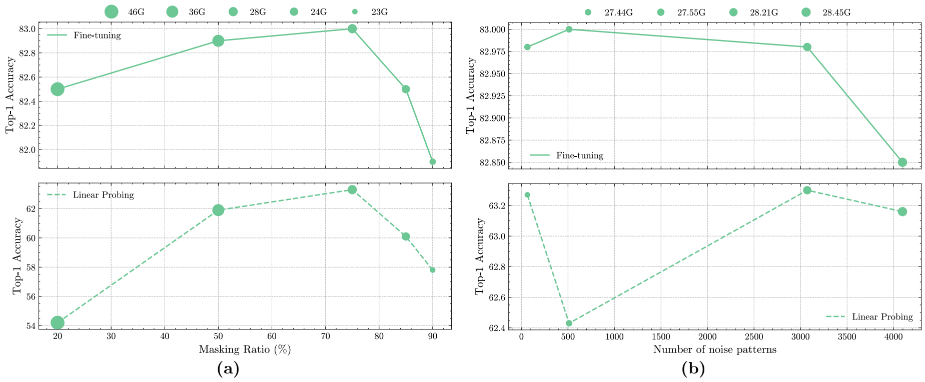

Fig. 1: Comparative analysis of Top-1 accuracy variations about different experimental settings. (a) Illustrates the Top-1 accuracy against five distinct masking ratios, showcasing how varying levels of masking influence the model's performance and its memory consumption per GPU during pre-training, represented by the size of the green circles. (b) Depicts the Top-1 accuracy across different amounts of noise patterns, illustrating their effect in accuracy. The results across both sets of plots indicate no overwhelming values for either masking ratios or noise patterns that consistently maximize Top-1 accuracy, suggesting a nuanced influence of these parameters on model performance.

2 Varying Masking Ratio

Fig. 1 (a) and Tab. 1 (a) illustrate the impact of varying masking ratios on the performance of our ColorMAE-G with ViT-B as the backbone. We conduct pre-training for 300 epochs across different masking ratios: 20%, 50%, 75%, 85%, and 90%. Similarly to MAE [1], we find that 75% works well for both fine-tuning (Fig. 1(a) top) and linear probing (Fig. 1(a) bottom). In the figure, y-axes correspond to ImageNet-1K validation accuracy Top-1 (%), and the memory per GPU used during pre-training for each masking ratio is shown at the top. Tab. 1 (a) shows more detailed information about the experiments, including the Top-1 and Top-5 accuracy metrics. In both the figure and the table, we show the memory used during pre-training. As observed, while a 50% masking ratio shows competitive fine-tuning performance, employing a 75% ratio is more memory-efficient and yields superior linear probing results.

3 Varying number of patterns

Similarly, Fig. 1 (b) and Tab. 1 (b) present the performance of ColorMAE-G with ViT-B when varying the number of noise patterns utilized during pre-training to produce the binary masks with Algorithm 1. We perform supervised training to evaluate the learned representations with end-to-end fine-tuning (Fig. 1 (b) top) and linear probing (Fig. 1 (b) bottom) and report the ImageNet-1K validation accuracy Top-1 (%). Tab. 1 (b) shows more detailed information about the experiments, including memory usage per GPU during pre-training and Top-1 and

Table 1: Quantitative analysis on ImageNet-1K classification tasks. We pre-train the MAE or ColorMAE-G with 300 epochs on ImageNet-1K, and report the Top-1 and Top-5 accuracy under the fully supervised fine-tuning and linear probing.

Masking Ratio	Fine-tuning		Linear Prob.		Memory
	Top-1	Top-5	Top-1	Top-5	GB
20	82.51	96.16	54.18	77.42	46.26
50	82.97	96.50	61.92	83.58	36.40
75	82.98	96.43	63.30	84.82	28.21
85	82.54	96.26	60.12	82.53	24.82
90	81.89	96.01	57.84	80.83	23.18

(a) ColorMAE-G with different masking ratios: Similar to MAE, we find 75% masking ratio leads to the best downstream performance.

Number of Patterns	Fine-tuning		Linear Prob.		Memory
	Top-1	Top-5	Top-1	Top-5	GB
64	82.98	96.40	63.27	84.51	27.44
512	83.00	96.43	62.43	84.08	27.55
3072	82.98	96.43	63.30	84.82	28.21
4096	82.85	96.43	63.16	84.33	28.45

(b) ColorMAE-G with different noise patterns: We do not find significant performance variations when changing the pattern numbers.

Random Patterns	Fine-tuning		Linear Probing	
	Top-1	Top-5	Top-1	Top-5
Uniform	82.82	96.32	60.70	82.66
White	82.63	96.33	60.65	82.41

(c) MAE with different noises for mask sampling: White noise can achieve performance comparable to that of uniform noise, which is the default noise used in MAE.

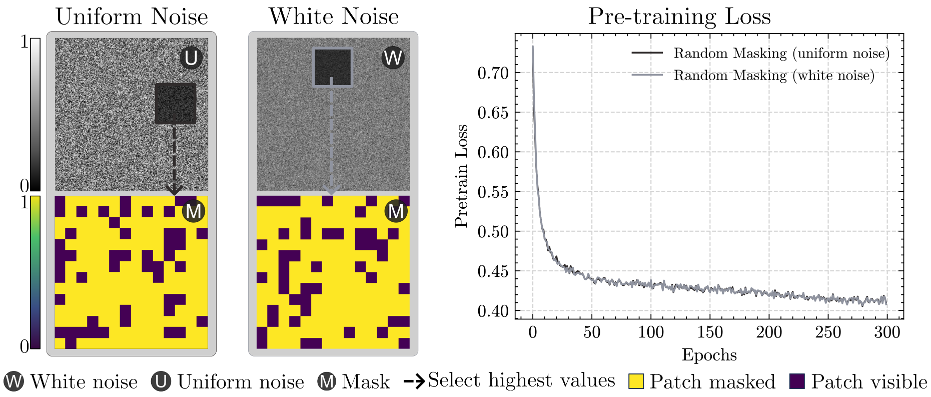

Fig. 2: Comparison of two randomly generated masks Ⓜ using uniform noise Ⓤ (first column) and white noise Ⓦ (second column). Despite their distinct noise origins, both result in remarkably similar random masks, as shown on the left and exhibit analogous pre-training loss, as illustrated on the right.

Top-5 accuracy metrics. The figure and the table show that naturally using more patterns directly increases memory usage. However, significant performance variations are not observed when varying the number of noise patterns. For example, employing only 64 green noise patterns of 256×256 spatial dimensions generally yields satisfactory results. It consumes nearly the same amount of memory as the original MAE [1] (27.44 GB) in our experiments. We decided to use 3072 noise patterns in all our experiments in the main paper as it provides the best balance between fine-tuning and linear probing accuracy. Additionally, the required memory of 28.21 GB was within the capacity limits of the GPUs utilized. However, fewer patterns can be employed without significant performance impact.

4 White Noise *vs*. Uniform Noise

Although white noise and uniform noise have different statistical properties, they can generate similar binary random masks, as illustrated in Fig. 2; see Ⓜ. In our experiments, no significant differences were observed in the pre-training loss, fine-tuning, or linear probing performance, as indicated in Tab. 1 (c). Therefore, within the context of MAE, using either uniform or white noise for generating binary masks yields comparable outcomes. Additionally, filtering operations can be applied to both white and uniform noise to produce various noise colors, generating analogous noise color patterns in our experiments.

5 Reconstruction Visualizations

In Fig. 3, we present additional visualizations of ImageNet validation images reconstructed using our ColorMAE, which was pre-trained with our four distinct types of masks: Blue, Green, Purple, and Red. For comparative analysis, we also include results from Random Masking, as utilized in the original MAE [1]. The visualizations highlight that our proposed masks generate unique patterns, leading ColorMAE to learn feature representations in distinct manners.

6 Self-attention Maps Visualizations

Fig. 4 presents supplementary self-attention map results for ColorMAE pre-trained with our green masking approach (ColorMAE-G), in comparison with the original MAE employing random masking. The attention maps are showcased on images sourced from the ImageNet-1K, Microsoft COCO, and ADE20K datasets, providing a comprehensive visual exploration across diverse data sets.

7 CAM Visualizations

Fig. 5 provides additional Class Activation Maps (CAM) of ViT-B after fine-tuning the model on image classification tasks, initialized with ColorMAE-G pre-training or conventional MAE pre-training. We employ EigenCAM [3] to highlight the model focus areas. As we can observe, the CAMs demonstrate that ColorMAE-G shows higher attention response on the foreground objects or peoples compared to MAE, suggesting our proposed approach can learn more discriminative visual representations.

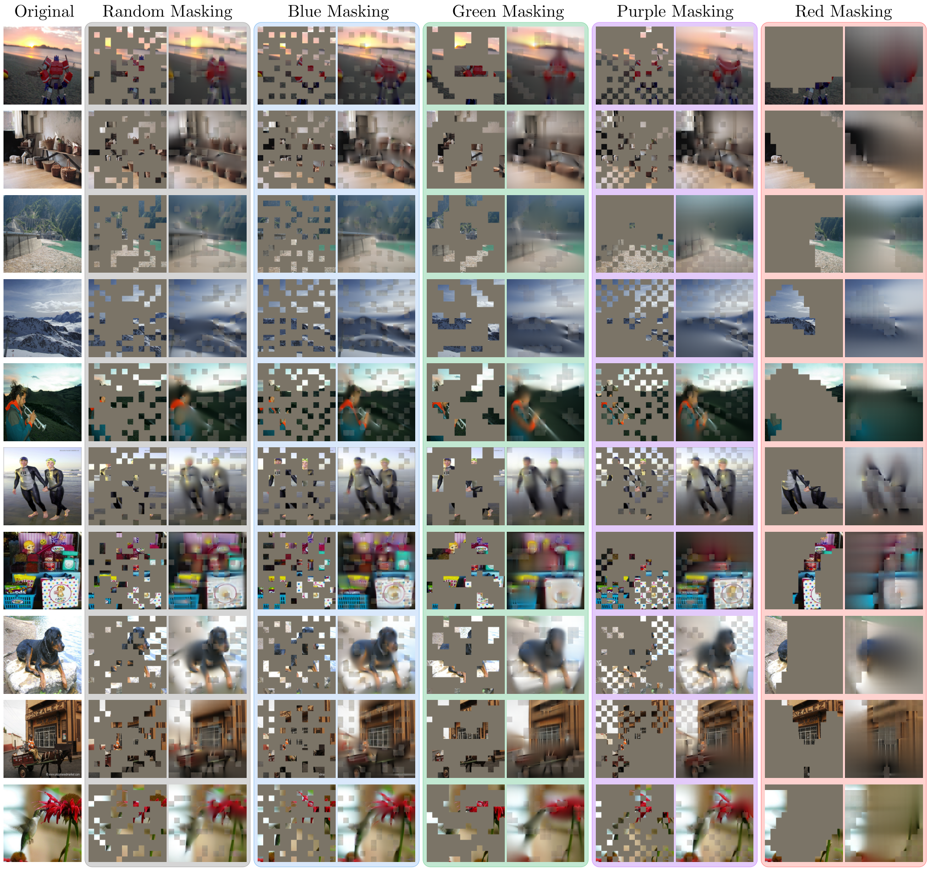

Fig. 3: Reconstruction results on ImageNet validation images from MAE pre-trained during 300 epochs with random masking and our four generated masks: Blue, Green, Purple, and Red.

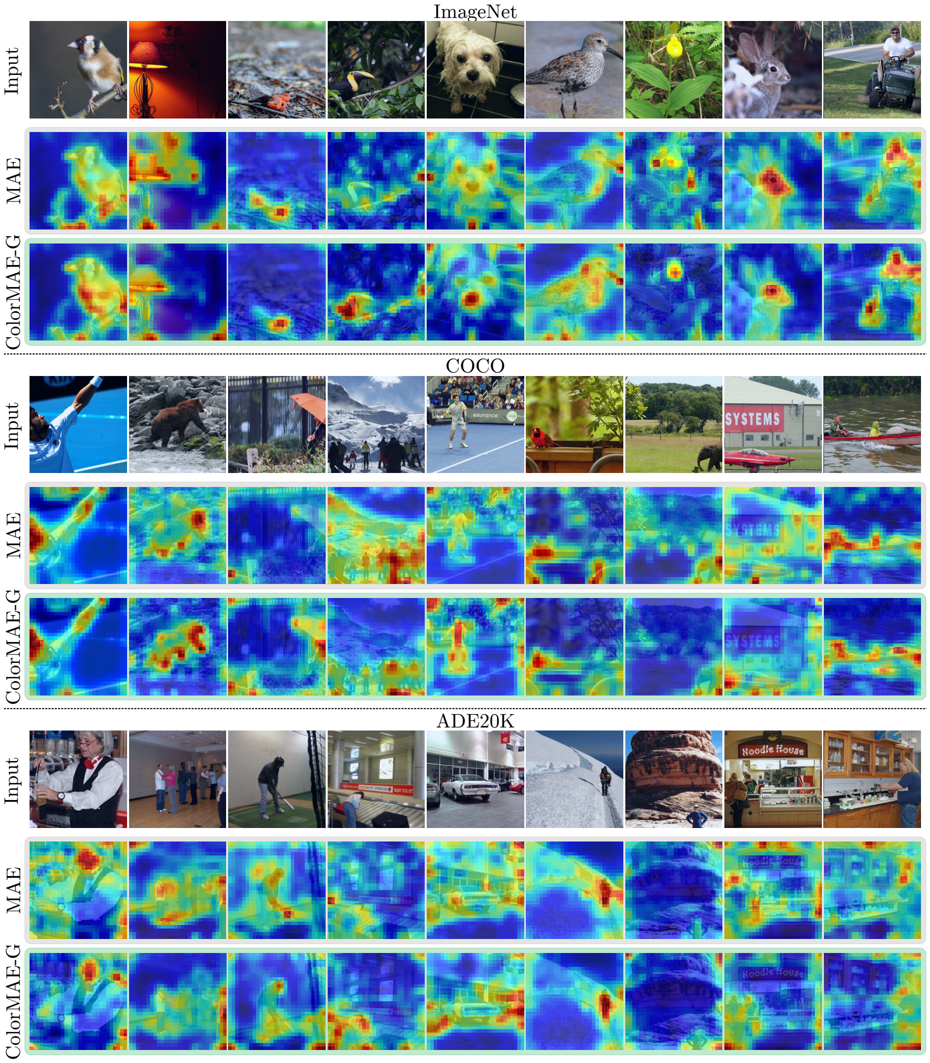

Fig. 4: Self-attention of the [CLS] tokens averaged across the heads of the last layer in MAE pre-trained using random masking and our proposed Green masking approach (ColorMAE-G). We show attention maps on images from Imagenet-1K [4](1st block), Microsoft COCO [2](2nd block) and ADE20K [5](3rd block) datasets.

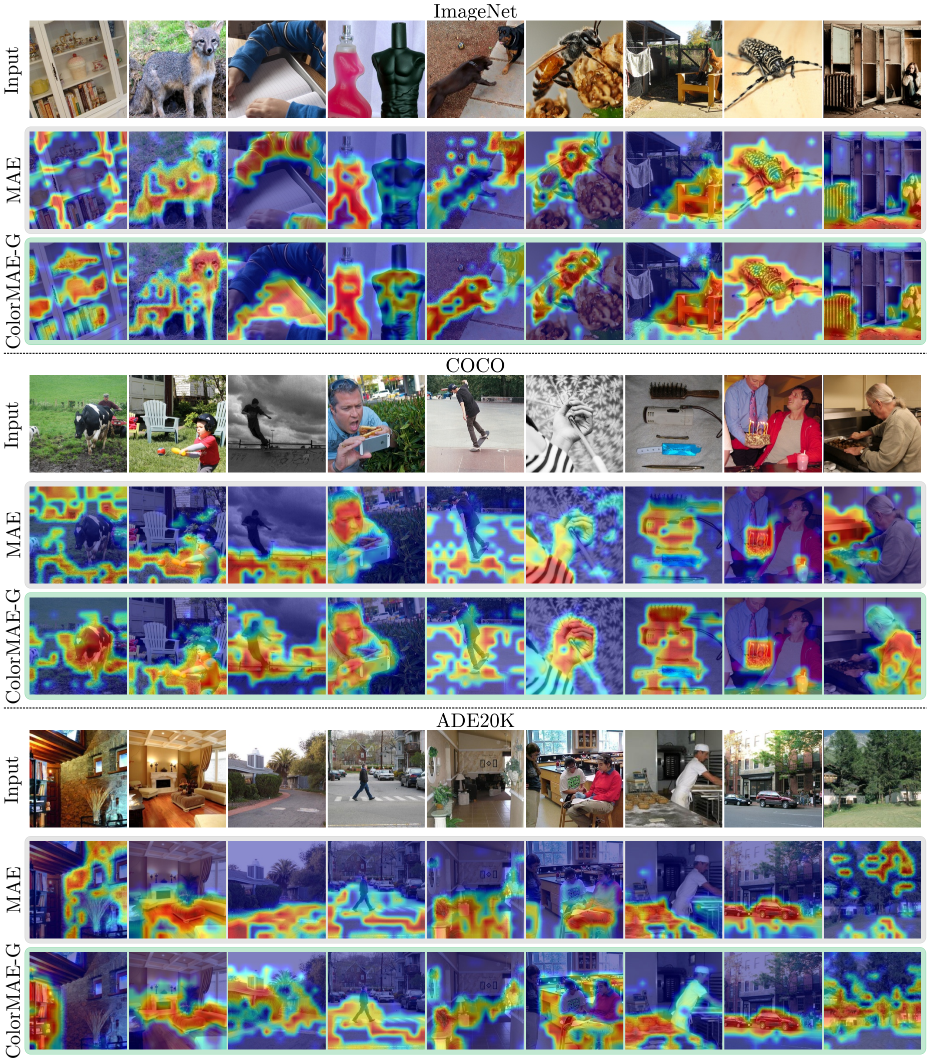

Fig. 5: Comparative visualization of Class Activation Maps (CAM) generated with EigenCAM [3] for ViT-B. We show CAM maps of the ViT-B pre-trained with MAE (second row of each block) and our ColorMAE-G (third row of each block) on images from ImageNet-1K (1st block), Microsoft COCO (2nd block), and ADE20K (3rd block) datasets.

References

1. 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)
2. 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)
3. 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)
4. 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)
5. 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)