This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

From Linearity to Non-Linearity: How Masked Autoencoders Capture Spatial Correlations

Anthony Bisulco*
University of Pennsylvania
3317 Chestnut St, Philadelphia PA
abisulco@seas.upenn.edu

Rahul Ramesh*
University of Pennsylvania
3317 Chestnut St, Philadelphia PA
rahulram@seas.upenn.edu

Randall Balliestro
Brown University
115 Waterman St, Providence, RI
randall_balestriero@brown.edu

Pratik Chaudhari
University of Pennsylvania
3317 Chestnut St, Philadelphia PA
pratikac@seas.upenn.edu

Abstract

Masked Autoencoders (MAEs) have emerged as a powerful pretraining technique for vision foundation models. Despite their effectiveness, they require extensive hyperparameter tuning (masking ratio, patch size, encoder/decoder layers) when applied to novel datasets. While prior theoretical works have analyzed MAEs in terms of their attention patterns and hierarchical latent variable models, the connection between MAE hyperparameters and performance on downstream tasks is relatively unexplored. This work investigates how MAEs learn spatial correlations in the input image. We analytically derive the features learned by a linear MAE and show that masking ratio and patch size can be used to select for features that capture short- and long-range spatial correlations. We extend this analysis to non-linear MAEs to show that MAE representations adapt to spatial correlations in the dataset, beyond second-order statistics. Finally, we discuss some insights on how to select MAE hyper-parameters in practice.

1. Introduction

Masked autoencoders (MAEs) are a powerful pretraining technique for images [20, 54] and video [17, 46, 48] playing a key role in pretraining for foundational models such as Segment Anything [28], InternVideo [49], EVA [16], and Unified IO [35]. Despite their success, MAEs are not yet well understood—key design choices such as masking ratio, patch size, and model depth heavily rely on heuristics.

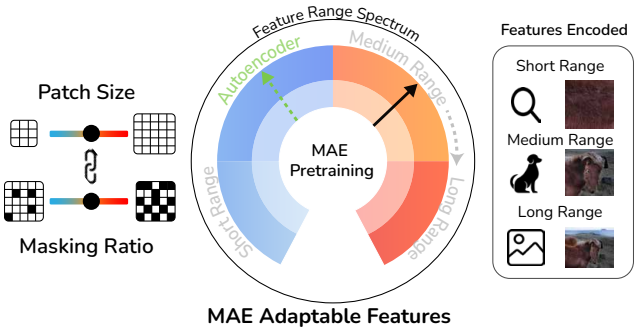


Figure 1. MAEs (Black Arrow) can adapt the spatial scale of features they encode through two key hyperparameters: patch size and masking ratio. Generally, increasing either of these hyperparameters yields features with a broader spatial extent. Selecting these hyperparameters allows MAEs to be used for datasets with different spatial scales. In contrast, standard autoencoders (Green Arrow) learn short-range spatial features. On the right side of the figure, we show an example of feature spatial scales ranging from local textures to objects to global scenes.

In this work, we aim to better understand MAEs by analyzing their underlying mechanisms and their role in representation learning.

A key challenge in applying MAEs to new datasets or modalities is the need for extensive cross-validation to tune hyperparameters such as patch size and masking ratio. Suboptimal choices can lead to degraded task performance and unnecessary computational cost. Exhaustive search over these parameters is often impractical, as MAEs' long training times make full cross-validation prohibitively expensive for large datasets or high-dimensional inputs. As a step towards addressing this challenge, we aim to study how masking ratio and patch size affect the representations learned by

*Equal Contribution

AB supported by NSF FRR 2220868, NSF IIS-RI 2212433, ARO MURI W911NF-20-1-0080, and ONR N00014-22-1-2677, and AB, RR, and PC supported by NSF IIS-2145164, CCF-2212519, ONR N00014-22-1-2255, and DSO Laboratories, Singapore.

MAEs (Fig. 1).

Our central hypothesis is that *MAEs learn spatial correlations in the input image*, with the masking ratio and patch size controlling the spatial scale of these correlations. We argue that MAEs introduce a spatial locality bias in ViT (Vision Transformer) architectures that inherently lack it, although some efforts have attempted to explicitly incorporate such biases [14, 34]. We make the following key contributions:

1. We derive an analytical expression for the encoder and decoder of a linear MAE. We show how MAEs can learn features that capture short- or long- range correlations by controlling the masking ratio and patch size.
2. We show that MAEs can be viewed as learning an adaptive basis for denoising, progressively shifting from localized to global feature representations over the course of training.
3. We provide insights for practitioners on how to select hyperparameters such as masking ratio, patch size, and number of encoder/decoder layers in MAEs.

2. Linear Masked Autoencoders

Reconstruction is a popular objective for self-supervised learning, but it often fails to learn useful features for downstream perception tasks [6]. Masked reconstruction [20, 54] has emerged as a popular choice for pretraining deep networks. In this section, we study how reconstruction differs from masked reconstruction and why the latter is useful for downstream perception tasks.



Figure 2. A $d = 8$ dimensional input can be divided up into 4 patches, each of size $p = 2$.

We first consider a linear MAE and analyze the solutions obtained for different masking ratios and patch sizes. Consider input data $X \in \mathbb{R}^{n\times d}$ where n is the number of samples, and each sample is a real-valued vector in d dimensions. Consider a linear encoder with weights $A \in \mathbb{R}^{d\times k}$ and a linear decoder with weights $B \in \mathbb{R}^{k\times d}$—the network has no non-linearities. The encoder projects the input from d input dimensions to k latent dimensions, and the decoder projects the representation back to d dimensions. A linear MAE minimizes the objective

$$\ell_m(A, B) = \mathbb{E}_R \left[\|X - (R \odot X)AB\|^2 \right], \qquad (1)$$

where $R \in \{0, 1\}^{n\times d}$ is a random variable denoting a mask and $\odot$ is the element-wise product. We can arrange our d-dimensional input vector into patches, each of size p (see Fig. 2). This is similar to how images are divided into patches in ViTs [13]. Patches are either fully masked or fully visible; this means $R_{ki} = R_{kj}$ if i and j belong to the same patch for any image with index k. Each patch is masked according to a Bernoulli random variable with probability m.[1]

Lemma 1 computes the expectation in Eq. (1) in closed form to reduce the linear MAE objective to

$$\ell_m(A, B) = \underbrace{\|X - (1-m)XAB\|^2}_{\text{reconstruction}} + m(1-m)\underbrace{\|GAB\|^2}_{\text{regularizer}} \qquad (2)$$

where $G \in \mathbb{R}^{d\times d}$ satisfies $G^\top G = \text{blkdiag}_p(X^\top X)$. The block-diagonal matrix $\text{blkdiag}_p(X^\top X)$ consists of block diagonal entries of size $p \times p$, with the off-diagonal blocks set to zero. The MAE objective in Eq. (2) consists of two loss terms: the first term is like the reconstruction loss of an autoencoder that encourages $X \approx (1 - m)XAB$, the second term $\|GAB\|^2$ acts as a regularizer tuned to the data via G. The masking ratio m controls the strength of the regularizer while the patch size controls the block-diagonal structure of the regularizer. Note that $m = 0$ recovers the objective for a non-masked autoencoder (AE). The regularizer term here, therefore, is the "bias" of an MAE. It forces the representation of an MAE to deviate from that of an AE.

2.1. Characterizing the minima of MAEs

We next analyze the encoder obtained in a linear MAE using Eq. (2). The classic work of Baldi and Hornik [4] shows that AEs are closely related to principal component analysis (PCA) [4]. AE's encoder at any global minimum extracts the top-k principal components of the data. Linear AEs, therefore, discover features that best explain the variance in the input data. The training objective of a linear AE does not have local minima. All global minima are of the form $A = UC^{-1}$, $B = CU^\top$, where the columns of $U \in \mathbb{R}^{d\times k}$ are the top-k eigenvectors of $X^\top X$, and $C \in \mathbb{R}^{k\times k}$ is any invertible matrix.

We can perform a similar analysis for linear MAEs to derive an analytical expression for the optimal encoder and decoder and characterize the set of all critical points of this objective. Note that while there are no non-linearities in a linear MAE, the training objective is still non-convex.

Theorem 1 *Let $V = (1 - m)X^\top X + m\,\text{blkdiag}_p(X^\top X)$ and let the columns of U_k denote the top-k eigenvectors of $X^\top XV^{-1}X^\top X$. The objective in Eq. (1) is minimized when the decoder*

$$B = CU_k \qquad (3)$$

and the encoder

$$A = V^{-1}X^\top XB^\top(BB^\top)^{-1}C^{-1}, \qquad (4)$$

[1] Typical implementations of MAEs exactly mask a fraction m of the patches instead of using a random variable to determine if a patch should be masked. Typically, MAEs also reconstruct only the unmasked patches at the output of the decoder, instead of employing an objective that reconstructs all patches, as we have done here.

where C is any invertible matrix of size $k \times k$.

Every critical point of the MAE objective is a subset of k eigenvectors of $X^\top X V^{-1} X^\top X$ (from Lemma 2), i.e., there are exponentially many critical points. The product of the encoder and decoder for an MAE comprises two parts. The first part $V^{-1} X^\top X$ whitens the data using a weighted mixture of $X^\top X$ and $\text{blkdiag}(X^\top X)$. The second part $B^\top (BB^\top)^{-1} B$, projects the data onto the column space of B.

Remark 1 If the input correlation matrix $X^\top X$ is itself block diagonal, i.e., spatial correlations in the input data are non-zero within a patch of size p, then $V = (1-m)X^\top X + m\,\text{blkdiag}_p(X^\top X) = X^\top X$, which recovers the regular autoencoder solution from Theorem 1, with a patch size of p. In this case, a linear autoencoder and a masked autoencoder have identical minima.

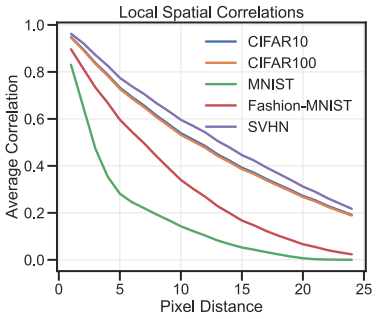


Figure 3. Spatial correlations between pixels as a function of the distance between them for different datasets. Correlation between pixels is inversely correlated to the distance for all 5 datasets, i.e., images have stronger local correlations.

Understanding the differences between an MAE and an AE using an Ising model. Images exhibit strong local correlations (Fig. 3) and the strength of the correlations between pixels is inversely proportional to the distance between them [25]. This structure can be quantified using the spatial autocorrelation:

$$R(\Delta x, \Delta y) = \frac{1}{MN} \sum_x \sum_y f(x,y) f(x+\Delta x, y+\Delta y) \quad (5)$$

which measures how a function $f(x,y)$ pixel values correlate across spatial distances $(\Delta x, \Delta y)$, for a 2D pixel grid of size (M, N). MAEs can exploit these inherent correlations in natural image data for masked patch reconstruction. To emulate a part of this structure, we consider data drawn from an Ising model. This model exhibits strong local correlations controlled by the coupling constant J. Under the Ising model, the input $x \in \{-1, 1\}^d$ has probability

$$p(x) \propto \exp\left(J x_1 x_d + J \sum_{i=1}^{d-1} x_i x_{i+1} \right). \quad (6)$$

We fit a linear AE and a linear MAE on samples from this distribution using the objective in Eq. (1) for the MAE. We approximate the correlation between x_i and x_j by $\langle x_i, x_j \rangle = \tanh(J)^r$, which is its value in the asymptotic limit of $d \to \infty$, where $r = \min(|i-j|, d-|i-j|)$ is the distance between x_i and x_j [23]. We use this to approximate $X^\top X$, which can be substituted in the analytical expressions in Theorem 1.

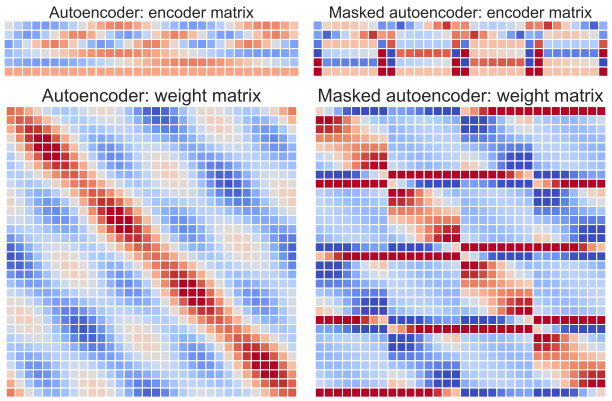


Figure 4. We plot the product of the encoder and decoder matrices (bottom) and the encoder weight matrix (top) for an AE (left) and MAE (right). Red indicates weights with higher magnitude, while blue indicates weights with lower magnitude. The plots show that MAEs learn an encoder that projects the data onto a basis that is different from a standard autoencoder. MAEs prioritizes features at the boundary of the patch since it is most correlated to the other patches and hence most useful for predicting them.

In Fig. 4, we plot the encoder weights (A) and the product of the encoder and decoder weights (AB) for a AE and MAE trained on the Ising model with $d = 32$ dimensions and coupling constant $J = 2$. Both the MAE and AE consider an encoder that projects data from 32 to 6 dimensions. For the MAE, we consider a patch size of 8 and a masking ratio of $m = 0.5$. For the AE (Fig. 4 left), the matrix AB has rank 6 (which is the dimensionality of the feature space). The product AB should be close to identity to minimize the autoencoder objective, and this is indeed evident in the experiment. In Fig. 4, the AE encoder learns sinusoidal features of varying frequencies, similar to those learned by linear autoencoders on natural images [19].

What is remarkable in this simple experiment is that an MAE learns a different encoder, one that clearly prioritizes input dimensions at the boundary of the patches. Since data has strong local correlations, the boundary of a patch has the largest correlation to nearby patches and is therefore most useful for reconstructing nearby patches. **In summary, MAEs select features that are redundantly present across patches while autoencoders pick features that best explain the variance in the data.** MAEs are perhaps successful on downstream tasks because many per-

ception tasks are redundant functions of the input [40], and MAEs find such redundant features.

2.2. Features of a linear MAE for natural images

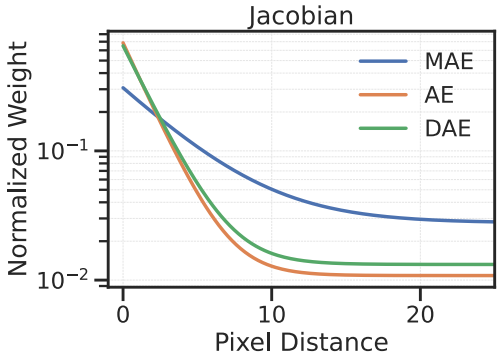


Figure 5. An exponential fit to the magnitude of the entries of the input-output Jacobian as a function of distance from the output pixel from the input pixel of different models (averaged over input and output pixels). Results are shown for three methods: MAEs with 80% masking ratio and patch size $p = 2$, standard AEs, and Denoising Autoencoders (DAEs) with noise level $\sigma = 0.2$. Experiments are conducted on CIFAR-10 with a latent dimension of 128. In general, MAEs integrate spatial information from distant patches, as opposed to an AE or a DAE.

We next train on natural images from CIFAR-10 and inspect how linear MAEs, linear denoising autoencoders (DAEs), and linear AEs encode different types of features. Unlike MAEs, which perform masking, DAEs add Gaussian random noise to the input [47]. For linear encoders and decoders, DAEs (Eq. (7)), denoise additive Gaussian noise with variance σ^2, and are equivalent to autoencoders with an ℓ_2 regularization on the weights (AB).

$$\ell_d(A, B) = \mathbb{E}_{L\sim\mathcal{N}(0,\sigma^2)}\left[\|X - (X + L)AB\|_2^2\right] \quad (7)$$

For AEs, DAEs and MAEs, we compare the average influence that an input pixel has on an output pixel as a function of the distance between the two. To compute this influence, we consider the Jacobian of the output with respect to the input. For example, for our linear models, the quantity $|(AB)_{ij}|$ determines the influence of input pixel i, towards reconstructing target pixel j. We compute the average magnitude of the normalized absolute weight of the entries of the Jacobian as a function of the distance of the output pixel from the input pixel for linear AEs, DAEs and MAEs trained on CIFAR-10 in Fig. 5. We find that AEs learn more localized kernels and DAEs are slightly less localized than AEs. On the other hand, MAEs integrate spatially distant information. This is because the MAE objective encourages the use of other patches to reconstruct a patch.

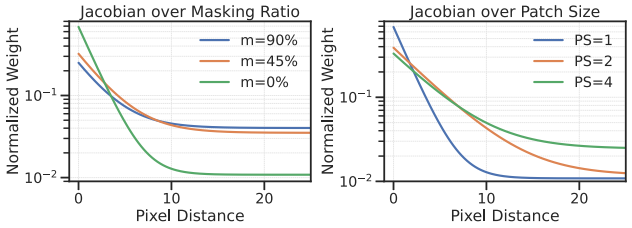


Figure 6. An exponential fit to the magnitude of the entries of the input-output Jacobian as a function of distance from the output pixel from the input pixel of different models (averaged over input and output pixels) for different masking ratios and patch sizes. Experiments were conducted for a linear MAE with latent dimension 128 on CIFAR-10. As the masking ratio increases, more and more non-local information is used for reconstruction in a (linear) MAE. Similarly, the reconstruction relies more on nonlocal information as the patch size increases.

Next, we investigate how MAE hyperparameters—masking ratio m and patch size p — affect the types of learned features. Mathematically, we can see this influence from Eq. (4), where the term $V^{-1}X^\top X$ first performs a whitening transformation of the original data. As the masking ratio increases, this transformation effectively performs patch-wise whitening, making all intra-patch correlations identity and normalizing inter-patch correlations. Masking forces the model to incorporate non-local information beyond the immediate patch for reconstruction, as seen in Fig. 6. Our results reveal that a high masking ratio produces a more diffuse average Jacobian, whereas a low masking ratio yields a more localized one. We observe a similar relationship with patch size, aligning with our prediction in Eq. (4): larger patch sizes expand the integration of information from regions outside the patch. In both of these cases, increasing the masking ratio and patch size reduces the weight of intra-patch pixels.

2.3. What kinds of tasks benefit from MAEs compared to AEs?

Previously, we showed that MAEs learn different features than AEs and DAEs, as illustrated in Fig. 5, e.g., MAEs capture more spatially distant information while reducing the emphasis on intra-patch details. This raises a key question: what kinds of tasks benefit more from MAEs compared to AEs? We investigate this through two tasks: a synthetic Gabor feature prediction task requiring integration of spatially distant information, and a monocular depth prediction task using the Cityscapes [10] dataset.

Fig. 6 suggests that MAEs capture more spatially distant information as the masking ratio and patch size increase. To investigate this, we evaluate them on predicting Gabor features, a task that allows precise control over the spatial dependencies required for accurate prediction. Gabor func-

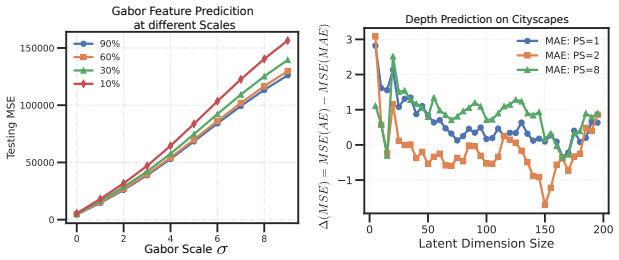


Figure 7. **Left:** Gabor feature prediction task using pretrained features from MAEs as a function of spatial scale (σ). In general, we see that a higher masking ratio, which forces the MAE to learn more non-local information, performs better on larger scales. **Right:** Reduction in MSE when an MAE is used for Cityscapes depth prediction as compared to an AE. For low-dimensional latents, MAEs outperform AEs on supervised depth prediction.

tions, denoted as $g(i, j)$ for spatial coordinates i, j, are localized harmonics modulated by a Gaussian window and are parameterized by frequency f, phase shift ϕ, Gaussian scale σ, and dilation γ:

$$g(i,j) = \exp\left(-\frac{i^2 + \gamma^2 j^2}{2\sigma^2}\right)\cos\left(2\pi i f + \phi\right). \quad (8)$$

Our target, $x_g(i, j)$, is the convolution of the Gabor function over the entire image $x_g(i, j) = (x * g)(i, j)$. The key advantage of using a Gabor feature prediction task is that we can adjust the parameter σ to control whether the task depends on local or long-range spatial interactions. This design enables a systematic evaluation of whether MAEs trained with higher masking ratios are better at capturing long-range dependencies. To analyze this effect, we trained an MAE with patch size ($p = 2$) and frequency values $f = [0.03, 0.06, 0.1]$, with fixed $\phi = 0$ and $\gamma = 1$, across various masking ratios (Fig. 7). For small σ values, different methods perform similarly. However, as σ increases, MAEs trained with high masking ratios outperform those with lower masking ratios. This aligns with our findings in Fig. 5, where MAEs integrate information from distant spatial patches.

While the Gabor prediction task provides a controlled setting to analyze MAEs, real-world perception tasks present greater complexity. To assess how MAEs perform in practical scenarios, we evaluate them on monocular depth prediction—a task that requires integrating spatial information to infer depth accurately. We trained MAEs on the Cityscapes dataset, downsampled to 32×32 resolution, and analyzed performance as a function of latent space dimensionality. For small latent dimensionalities, all patch sizes ($m = 0.8$) outperform an AE. This finding highlights how MAEs, through masking-based regularization, are useful for biasing representations for downstream perception tasks.

Linear MAEs are amenable to analysis but they come with their own limitations. (i) On datasets like CIFAR, ViTs use over-complete representations of the data, while our analysis is restricted to rank-deficient or full-rank encoders and decoders. (ii) Linear MAEs only capture 2nd-order correlations of the data. Higher order correlations in the data distribution can be used for better reconstruction, and possibly provide a better prior for downstream tasks. (iii) The network architecture for linear MAEs is limited to fully connected networks, which provides limited insight into designing ViTs for MAEs.

Section Summary

1. Linear MAEs learn a reconstruction kernel that biases feature learning towards spatially distant and correlated parts of the input by discouraging the use of intra-patch level information through the regularizer $||GAB||^2$.
2. Increasing the masking ratio or patch size forces MAE features to retain spatially distant information.
3. MAEs with high masking ratios perform better at tasks that require incorporating inter-patch level information.

3. Understanding Nonlinear MAEs

The objective for a non-linear MAE can be written as

$$\ell_m(\theta) = \mathbb{E}_R\left[\|X - f_\theta(g_\theta(R \odot X))\|^2\right], \quad (9)$$

where g_θ is the encoder and f_θ is the decoder, and we define: $f_\theta(g_\theta(R \odot X)) = h_\theta(R \odot X)$. For MAEs, the encoder and decoder are ViT blocks, where the encoder processes only the unmasked patches, and the decoder processes all patches. We could extend our analysis to non-linear MAEs using a linear approximation of MAEs (see Appendix B). However, these approximations give us little insight into how we should train MAEs using deep networks.

To analyze the encoder and decoder, we adopt the technique of Kadkhodaie et al. [27], which draws a connection between diffusion models and denoising. We can draw a similar parallel between MAEs and denoisers, except that the denoising task is set up using masking instead of corruption by additive Gaussian noise. Kadkhodaie et al. [27] argue that diffusion models learn an adaptive low-dimensional basis from data and we can identify this basis using techniques from Mohan et al. [37]. Extracting this basis requires the network to be a first-order homogeneous function, which ViTs are not—operations like layer-norm and attention are not homogeneous. To understand the basis learned by MAEs trained using ViTs, we approximate our denoiser $h_\theta(R \odot x)$ via a first-order Taylor series expansion around an input $\tilde{x} = R \odot x$, denoted by x_0. The

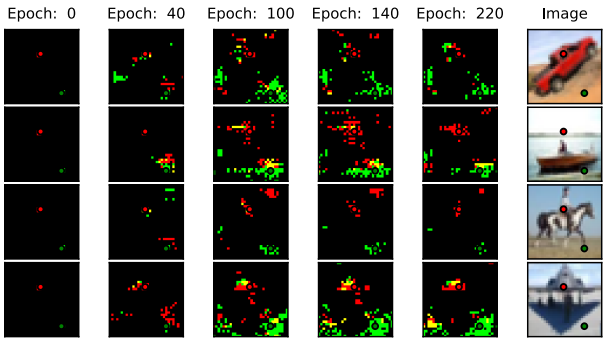


Figure 8. Visualization of the Jacobian of the output with respect to CIFAR-10 images, showing which input pixels contribute to reconstructing a given output pixel. We track this throughout training for the green and red pixel locations, showing their respective reconstruction kernels in corresponding colors. Each row presents a different image. In general, the Jacobian evolves throughout training, transitioning from highly localized to broader spatial dependencies, and adapts dynamically to each sample. The network is trained with a 80% masking ratio and a patch size of 2.

approximation can be written as:

$$h(\tilde{x}) \approx h(x_0) + \nabla_{\tilde{x}} h(\tilde{x})(\tilde{x} - x_0) = A(\tilde{x}) + b, \qquad (10)$$

where $A = \nabla_{\tilde{x}} h(\tilde{x})$ is the Jacobian of the reconstruction with respect to the inputs and $b = f(x_0) - \nabla_{\tilde{x}} h(\tilde{x}) x_0$ is the bias term. Assuming that the Jacobian does not change locally [5], $\nabla_{\tilde{x}} h(\tilde{x}) = \nabla_{x_0} h(x_0)$. We use this technique as a heuristic to analyze a nonlinear MAE, and thus interpret the encoder and decoder as learning an adaptive basis matrix $A = \nabla_{\tilde{x}} h(\tilde{x})$ for each data sample.

Fig. 8 illustrates the evolution of the learned basis throughout training for a MAE-trained ViT with a masking ratio of 80% and a patch size of 2. The figure highlights two different output pixels, circled in red and green, with corresponding colored input features identified by the Jacobian. First, unlike the linear MAEs discussed earlier, the basis used for reconstruction is adaptive and varies across images rather than being fixed. Second, during training, the basis transitions from being highly localized at the start to more diffuse, incorporating information from distant regions as training progresses. Third, while the linear model relies primarily on second-order correlations, ViTs leverage higher-order correlations to improve the performance on masked reconstruction. For instance, in the horse image, the reconstruction kernel adapts to capture correlations with the human's shirt, while for background pixels, it dynamically adjusts across similar regions. This higher-order correlation modeling mitigates the overly smoothed reconstructions typical of purely second-order models, demonstrating that our model leverages complex statistical dependencies.

Next, we examine the influence of masking ratio and patch size on the Jacobian matrix A. In Fig. 9, we demonstrate that increasing the patch size shifts the reconstruction from relying on local information to incorporating more global context. We characterize this using the spatial entropy for each output pixel. Given that the Jacobian of a single output pixel is denoted as $J \in \mathbb{R}^{HW}$, the spatial entropy is computed as:

$$S = -\sum_i \frac{|J_i|}{\|J\|_1} \log\left(\frac{|J_i|}{\|J\|_1}\right), \qquad (11)$$

where i is the ith component of J. We then construct a histogram of S for all output pixels. Generally, larger patch sizes result in higher spatial entropy, indicating broader use of information compared to smaller patches, while changes in masking ratio had a much weaker influence on distant spatial information. Beyond CIFAR, we also show similar results on ImageNet-64 in Fig. 13.

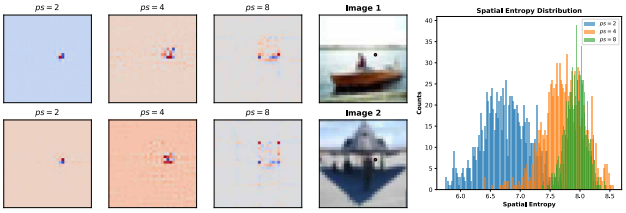


Figure 9. **Left:** Visualization of the Jacobian of the output with respect to CIFAR-10 images as we vary patch size (masking ratio fixed at $m = 0.6$) for the image in each row, for the denoted output pixel in red. As we increase the patch size, the weights have a less localized distribution. **Right:** Spatial Entropy distribution over all the output pixels for different patch sizes. In general, larger patch sizes result in distributions with higher spatial entropy.

Section Summary

1. Nonlinear MAEs develop an adaptive kernel for denoising that is specific to the test datum compared to the dataset-specific kernel of a linear MAE.
2. Non-linear MAEs grow over the course of training from a well-localized kernel at the start to a kernel with a broader spatial support at the end of training.
3. Similar to the linear MAE, increasing patch size increases the spatial extent of encoded features.

4. From theory to practice: How to train your MAE

Masked autoencoders require significantly more training epochs than other self-supervised learning methods to achieve competitive downstream performance. For example, on ImageNet, MAEs require 1,600 training epochs, whereas I-JEPA [1] achieves a similar task performance in just 300 epochs. Both pretraining and fine-tuning can be

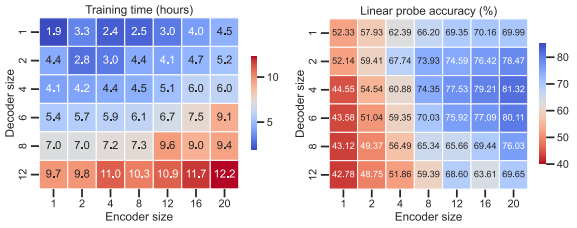


Figure 10. We plot (left) the training time and (right) linear probe accuracy for different combinations of the number of encoder and decoder layers. **MAEs are slow to train** with training time growing faster as we increase the size of the decoder. We find that training loss is not a good proxy for downstream classification accuracy Fig. 14. The **accuracy of the trained encoder continues to improve as we increase its size**. However, this is not true for the decoder: the optimal decoder contains 2-4 layers. We notice that the differences in accuracies reduce when we fine-tune the encoder for a 100 epochs Fig. 15

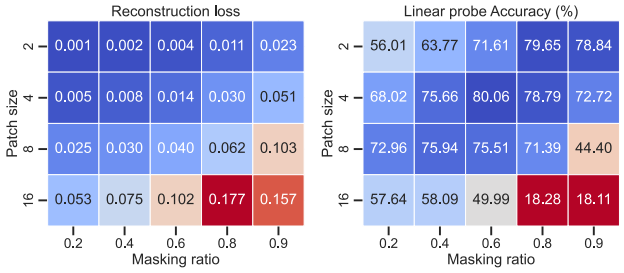


Figure 11. We plot the reconstruction error (left) and the linear probe accuracy (right) for different values of patch-size and masking ratio. Reconstruction loss decreases as we decrease the patch-size or decrease the masking ratio. However the linear probe accuracy has no clear trend with respect to the masking ratio and the optimal masking ratio depends on the patch-size. While smaller values of patch-size achieve better linear probe accuracies, they are also slower to train.

sensitive to the choice of hyperparameters, particularly with ViTs [45]. In this section, we discuss some empirical insights for MAE pretraining and fine-tuning based on our experiments on CIFAR-10.

4.1. Experimental details

We train MAEs using the architecture in He et al. [20]. We fix the embedding dimension d at 192, the number of attention heads in the encoder to 12 and the decoder to 16 in all our experiments. The Transformer blocks use the GeLU non-linearity [21] with the layer-norm before the attention and MLP blocks — also known as the pre layer-normalization transformer [56].

The MAE is trained for 2000 epochs using the AdamW optimizer with 50 epochs of warmup and a cosine-annealed learning rate schedule with an initial learning rate of 1.2×10^{-3}. The MAEs are trained with a batch-size of 4096 and weight-decay of 0.05. After pretraining, we discard the decoder and fine-tune the encoder. We fine-tune for only 100 epochs using a batch-size of 1024 with 5 epochs of warmup and an initial learning rate of 10^{-3} annealed to 10^{-5} using a cosine-annealed schedule. For MAE pretraining, we use random-resize crop and horizontal flips as the two augmentations, while for fine-tuning we use RandAugment [11].

We evaluate the accuracy on the downstream task using a linear-probe applied to the encoder. We find that trends obtained using a linear probe are nearly identical to trends obtained using fine-tuning (see Appendix D).

4.2. Results

How should we select the size of the encoder and decoder? Masked autoencoders typically have more encoder layers than decoder layers, but is this optimal? We investigate how the number of encoder and decoder layers affect the reconstruction error, training time and downstream classification accuracy. In Fig. 10, we train models on CIFAR-10 and plot the training time (in hours) as we vary the number of encoder and decoder layers. The training time increases at a slower rate when we increase the number of encoder layers compared to increasing the number of decoder layers since the decoder operates on a longer sequence of tokens — the decoder uses both masked and unmasked tokens while the encoder only uses unmasked tokens. Even for CIFAR-10, training can take as long as 6 to 10 hours on a single GPU.

Increasing the number of encoder and decoder layers decreases the reconstruction error (Fig. 14). However, a smaller reconstruction error does not imply higher linear-probe accuracies (Fig. 14) or fine-tuning accuracies (Fig. 15). **We find that the linear-probe accuracy continues to improve as we increase the size of the encoder. However, the optimal decoder has far fewer layers and even a 1-layer decoder has high accuracy after fine-tuning with the added benefit of reduced training time.** For example, a model pretrained with a 1-layer decoder achieves 95.26% accuracy after fine-tuning, which is only 0.30% worse than our best model Fig. 15.

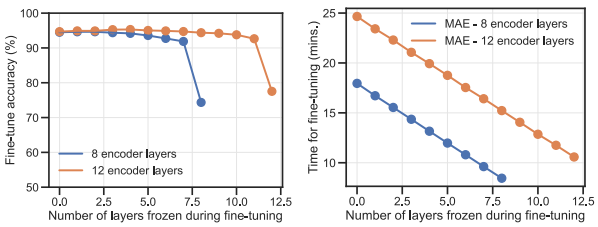


Figure 12. We fine-tune the model for 100 epochs after freezing the first k layers of the network and plot the accuracy against number of layers frozen during fine-tuning. We find that even if we freeze all but 1 Transformer block, we recover the accuracy compared to when no layers are frozen, i.e., the accuracy can be recovered while using a fraction of the training time and memory.

How do we select the masking ratio and patch-size? The masking ratio and patch-size control the basis learned by a linear MAE (Sec. 2). We see that these two hyperparameters also have a significant impact on MAEs trained using deep networks. In Fig. 11, we plot the reconstruction error and the linear-probe accuracy for different values of masking ratio and patch-size. **MAEs train faster if the patch-size is large or if the masking ratio is large** since a larger patch-size reduces the number of tokens by a quadratic factor and increasing the masking ratio decreases the number of tokens fed to the encoder. The training time increases from 4 hours to 10 hours as we decrease the masking ratio from 0.9 to 0.1.

Reconstruction improves as we decrease the masking ratio which is unsurprising — it is easier to reconstruct an image if we have more patches. Reconstruction error also reduces as the patch-size becomes smaller. A smaller patch-size reduces the average spatial distance to unmasked patches, which makes reconstruction easier due to strong local spatial correlations seen in images (Fig. 3). We find that a small patch-size and large masking ratio achieve the best linear probe accuracies. While smaller values of patch-size achieve better linear probe accuracies, an MAE with larger patch-size trains faster.

Can we fine-tune fewer layers to reduce GPU memory and training time? We investigate the effect of fine-tuning a subset of the layers of an MAE in Fig. 12. We find that **if we freeze all but one Transformer block, we achieve within 2% of the accuracy achieved from fine-tuning all the layers**, i.e., a similar accuracy can be achieved with a nearly 2x speedup (Fig. 15 (right)) and with less GPU memory. These results align with the practice of setting learning rates that exponentially decrease across the different layers of the encoder during fine-tuning [20].

Section Summary

1. MAEs achieve higher linear probe accuracy with more encoder layers and fewer decoder layers. On CIFAR-10, accuracy improves consistently with more layers, and a single-layer decoder reaches within 1% of the best fine-tuning accuracy while training 4× faster.
2. As the patch-size used to train the MAE increases, the masking ratio that achieves the highest linear-probe accuracy decreases.
3. On CIFAR-10, fine-tuning just the last transformer block is enough to achieve within 2% of the best accuracy.

5. Related Work

Theoretical work on understanding MAEs can be categorized into the following themes: architectural modifications, data modeling, connections to other self-supervised learning objectives and denoising. Previous works that study architectural variants find that masking alters the attention structure of ViTs, creating more local attention in early layers rather than global [8, 24, 38, 41, 55]. Raghu et al. [39] suggests that ViTs with local receptive fields generalize better to vision tasks. Kong et al. [29] hypothesizes that this learning process attempts to learn a hierarchical latent variable model, where masking enforces identifiability of coarser latent variables, though such a model is not definitively known to exist for natural image data. Another perspective interprets MAEs as a form of contrastive learning using only positive samples of unmasked patches [57]. Littwin et al. [32] challenges this hypothesis by demonstrating that MAE approaches learn a different set of features than contrastive methods. Beyond explaining the architectural bias induced by MAEs, researchers have also investigated how MAEs can perform masked image modeling tasks under extreme masking ratios. This challenge can be framed as a linear inverse problem, where measurements of unmasked pixels are used to recover masked pixels [26]. While compressive sensing offers one method to solve this inverse problem, it lacks the nonlinear encoder and prior over the data distribution present in MAEs [51]. For the nonlinear problem, the field of image denoising provides a probabilistic understanding of the unique properties induced in this denoising task [7, 36]. For additional related work see Appendix G

6. Conclusion

This work sheds light on how MAEs build strong foundational features. We showed that MAEs capture spatial correlations in input images. We model MAEs in a linear setting to demonstrate how masking ratio and patch size affect the integration of spatially distant features. We extend this study to nonlinear MAEs by analyzing the input-output Jacobian and provide practical training guidelines for practitioners.

While MAEs are adaptable, the sensitivity of their hyperparameters to dataset and task characteristics necessitates careful tuning. Understanding the connection between properties of the visual task and optimal MAE hyperparameters remains an important direction for future work. Overall, our characterization of MAEs as spatial correlation learners provides valuable insights for advancing their theoretical understanding.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael G. Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. *CVPR*, pages 15619–15629, 2023. 6

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, pages 1298–1312, 2022. 6

[3] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Loddon Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. *CVPR*, pages 24256–24265, 2023. 6

[4] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989. 2

[5] Randall Balestriero and Richard Baraniuk. A spline theory of deep learning. In *ICML*, 2018. 6

[6] Randall Balestriero and Yann LeCun. Learning by reconstruction produces uninformative features for perception. In *ICML*, 2024. 2

[7] José M. Bioucas-Dias and Mário A. T. Figueiredo. Multiplicative noise removal using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19:1720–1730, 2009. 8

[8] Shuhao Cao, Peng Xu, and David A. Clifton. How to understand masked autoencoders. *ArXiv*, abs/2202.03670, 2022. 8

[9] Minmin Chen, Kilian Q. Weinberger, Zhixiang Eddie Xu, and Fei Sha. Marginalizing stacked linear denoising autoencoders. *J. Mach. Learn. Res.*, 16:3849–3875, 2015. 6

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CVPR*, pages 3213–3223, 2016. 4

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 7

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019. 6

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

[14] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. ConViT: Improving vision transformers with soft convolutional inductive biases. In *ICML*, 2021. 2

[15] David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector J. Santos-Villalobos, M. V. Rohith, and Xinyu Li. Motion-guided masking for spatiotemporal representation learning. *ICCV*, 2023. 6

[16] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 1

[17] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *NeruIPS*, 35:35946–35958, 2022. 1

[18] Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, XuDong Wang, Adam Yala, Trevor Darrell, Alexei A Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *Transactions on Machine Learning Research*, 2025. 6

[19] Peter J. B. Hancock, Roland J. Baddeley, and Leslie S. Smith. The principal components of natural images. *Network: Computation In Neural Systems*, 3:61–70, 1992. 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. 1, 2, 7, 8, 3, 6

[21] Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. In *ICLR*, 2017. 7

[22] Vlad Hondru, Florinel Alin Croitoru, Shervin Minaee, Radu Tudor Ionescu, and Nicu Sebe. Masked image modeling: A survey, 2024. 6

[23] Kerson Huang. *Introduction to statistical physics*. Chapman and Hall/CRC, 2009. 3

[24] Yu Huang, Zixin Wen, Yuejie Chi, and Yingbin Liang. A theoretical analysis of self-supervised learning for vision transformers. *ICLR*, 2025. 8

[25] Aapo Hyvärinen, Jarmo Hurri, and Patrik O. Hoyer. Natural image statistics - a probabilistic approach to early computational vision. In *Computational Imaging and Vision*, 2009. 3

[26] Zahra Kadkhodaie and Eero P. Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *NeurIPS*, 2021. 8

[27] Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and St'ephane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. *ICLR*, 2024. 5

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *ICCV*, 2023. 1

[29] Lingjing Kong, Martin Q. Ma, Guan-Hong Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, and Kun Zhang. Understanding masked autoencoders via hierarchical latent variable models. *CVPR*, 2023. 8

[30] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *ICML*, 2019. 4

[31] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *NeurIPS*, 2022. 6

[32] Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *NeurIPS*, 2024. 8

[33] Hao Liu, Xinghua Jiang, Xin Li, Antai Guo, Yiqing Hu, Deqiang Jiang, and Bo Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pretraining. In *AIAA*, pages 1649–1656, 2023. 6

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2

[35] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. In *ICLR*, 2023. 1

[36] Peyman Milanfar and Mauricio Delbracio. Denoising: A powerful building-block for imaging, inverse problems, and machine learning. *ArXiv*, abs/2409.06219, 2024. 8

[37] Sreyas Mohan, Zahra Kadkhodaie, Eero P. Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. *ICLR*, 2020. 5

[38] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What Do Self-Supervised Vision Transformers Learn? In *ICLR*, 2023. 8

[39] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *NeurIPS*, 2021. 8

[40] Rahul Ramesh, Anthony Bisulco, Ronald W DiTullio, Linran Wei, Vijay Balasubramanian, Kostas Daniilidis, and Pratik Chaudhari. Many perception tasks are highly redundant functions of their input data. *COSYNE*, 2025. 4

[41] Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari S. Morcos. Understanding contrastive versus reconstructive self-supervised learning of vision transformers. In *NeurIPS*, 2022. 8

[42] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *CVPR*, 2023. 6

[43] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014. 6

[44] Harald Steck. Autoencoders that don't overfit towards the identity. In *NeurIPS*, 2020. 6

[45] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. 7

[46] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022. 1

[47] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 4, 6

[48] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. *CVPR*, 2023. 1

[49] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *ECCV 2024*, 2024. 1

[50] Chen Wei, Haoqi Fan, Saining Xie, Chaoxia Wu, Alan Loddon Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *CVPR*, 2022. 6

[51] John Wright and Y. Ma. High-dimensional data analysis with low-dimensional models. 2022. 8

[52] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. In *ICLR*, 2023. 6

[53] Johnathan Xie, Yoonho Lee, Annie S Chen, and Chelsea Finn. Self-guided masked autoencoders for domain-agnostic self-supervised learning. *ICLR*, 2024. 6

[54] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: a simple framework for masked image modeling. *CVPR*, 2022. 1, 2, 6

[55] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the Dark Secrets of Masked Image Modeling. In *CVPR*, 2023. 8

[56] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *ICML*, 2020. 7

[57] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *NeurIPS*, 2022. 8