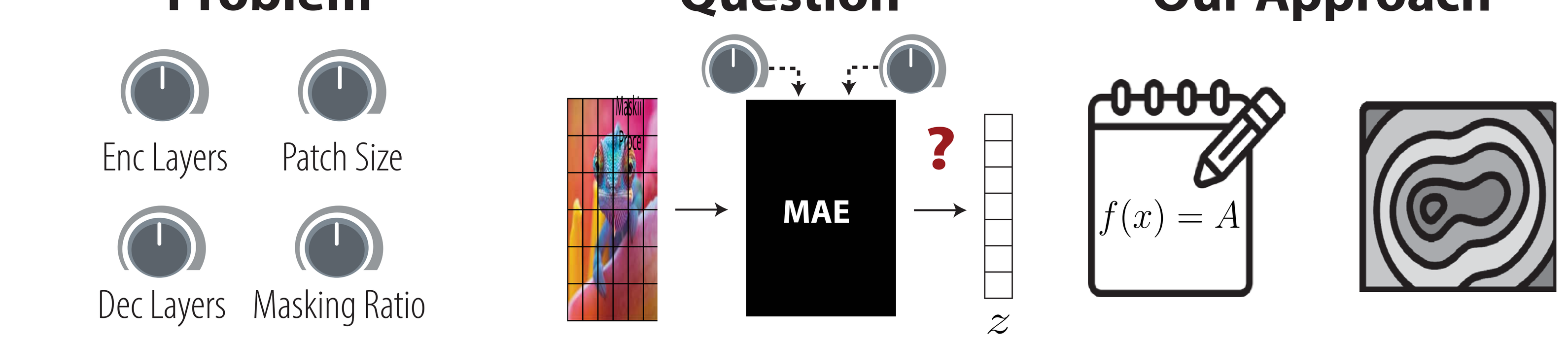
From Linearity to Non-Linearity: How Masked Autoencoders Capture Spatial Correlations

Anthony Bisulco*[1], Rahul Ramesh*[1], Randall Balestriero[2], Pratik Chaudhari[1]
[1]University of Pennsylvania, [2]Brown University, *Equal Contribution

ICCV HONOLULU HAWAII OCT 19-23, 2025

How should one choose MAE hyperparameters?

Problem | **Question** | **Our Approach**



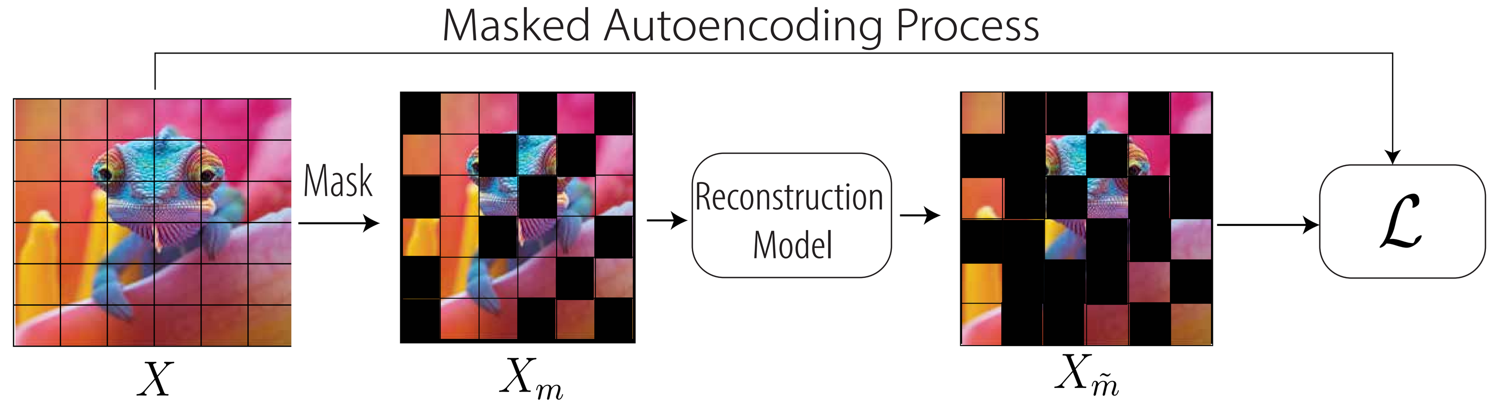
Strong task performance = extensive hyperparameter tuning

How do MAE hyperparameters control learned features?

Simplified Model

Experimental Validation

Masked Autoencoding Process

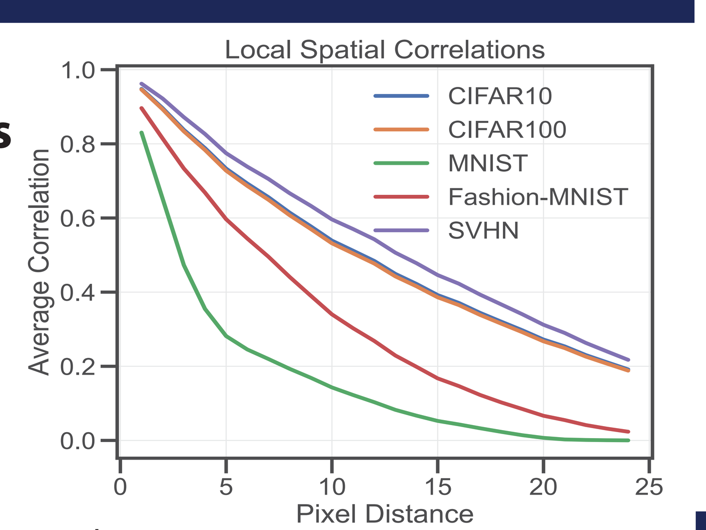


Background

- Natural images have a spatial structure: nearby pixels have **strong local correlations**
- This structure is captured in the spatial auto-correlation function:

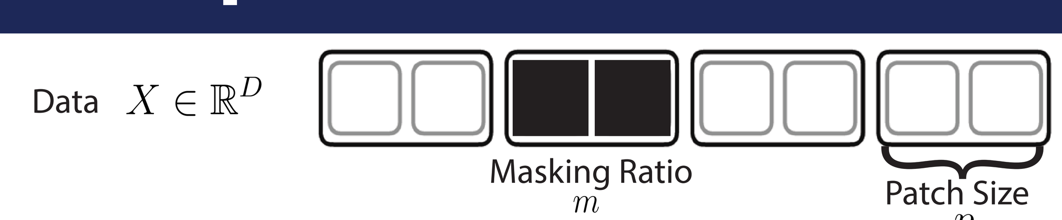
$$R(\Delta x) = 1/N \sum_x f(x) f(x+\Delta x)$$

- By tuning patch size and masking ratio, we control the spatial correlations the model can exploit



Takeaway: Statistical correlations in data provide regularities that MAEs can exploit to reconstruct masked regions

Simplified Model: Linear MAE



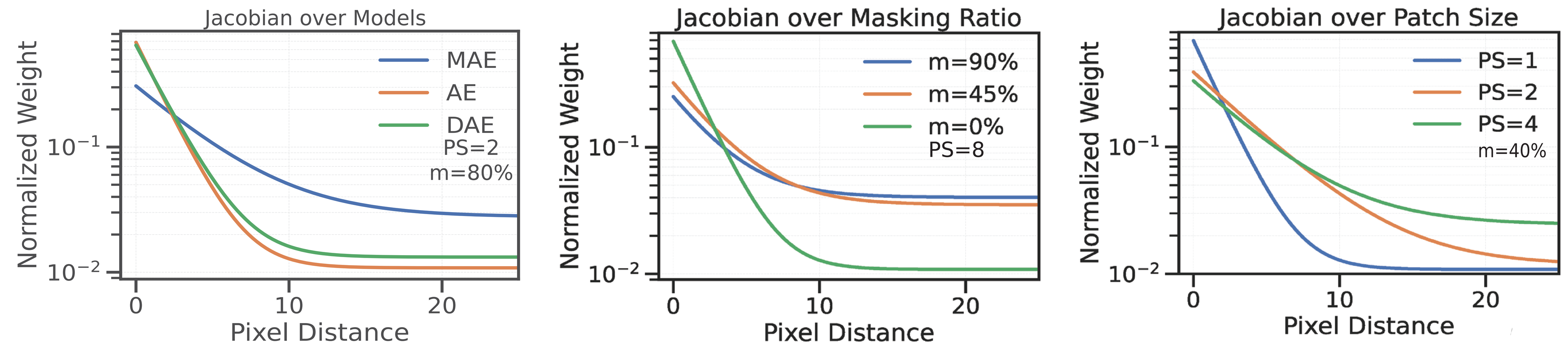
$$\mathcal{L}(A,B) = \mathbb{E}_R \|X - (R \odot X)AB\|^2 \xrightarrow{\text{Expectation}} \mathcal{L}(A,B) = \underbrace{\|X-(1-m)XAB\|^2}_{\text{reconstruction}} + m(1-m)\underbrace{\|GAB\|^2}_{\text{regularizer}}$$

Encoder: A, Decoder: B, Mask: R, Regularizer: $G^\top G = BlkDiag_p(X^\top X)$

- Marginalize linear MAE loss over masks → reconstruction + regularization (Bias of an MAE) terms → solve for closed form optimal solution
- The MAE bias makes it select features that are redundantly present across patches as opposed to an AE, which selects features that explain variance

Takeaway: Linear MAEs acts as a data-dependent regularized autoencoder, masking ratio sets the strength, patch size controls spatial structure

How does masking shape representations in linear MAEs?

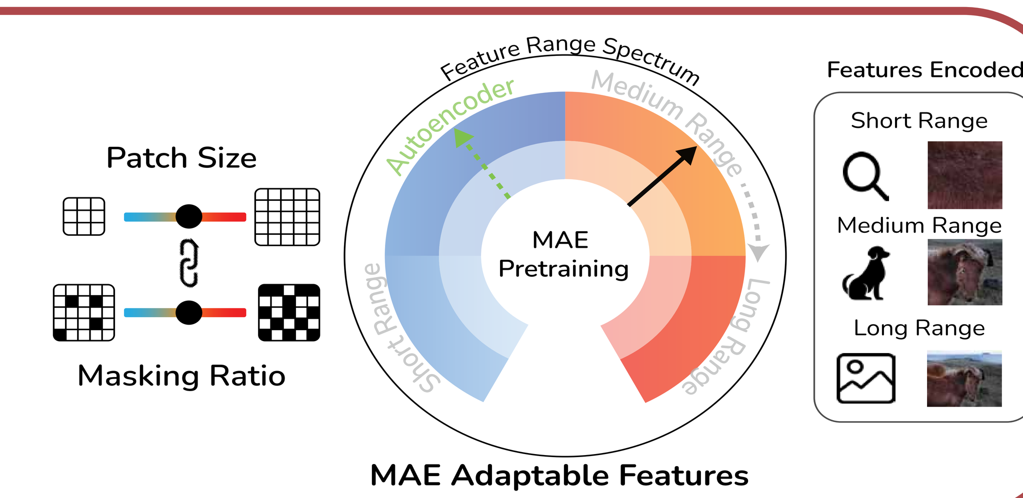


Jacobian $|(AB)_{ij}|$ determines the influence of input pixel i, towards reconstructing target pixel j. Magnitude averaged over inputs. Results for Linear MAE models on CIFAR-10; similar behavior on ImageNet.

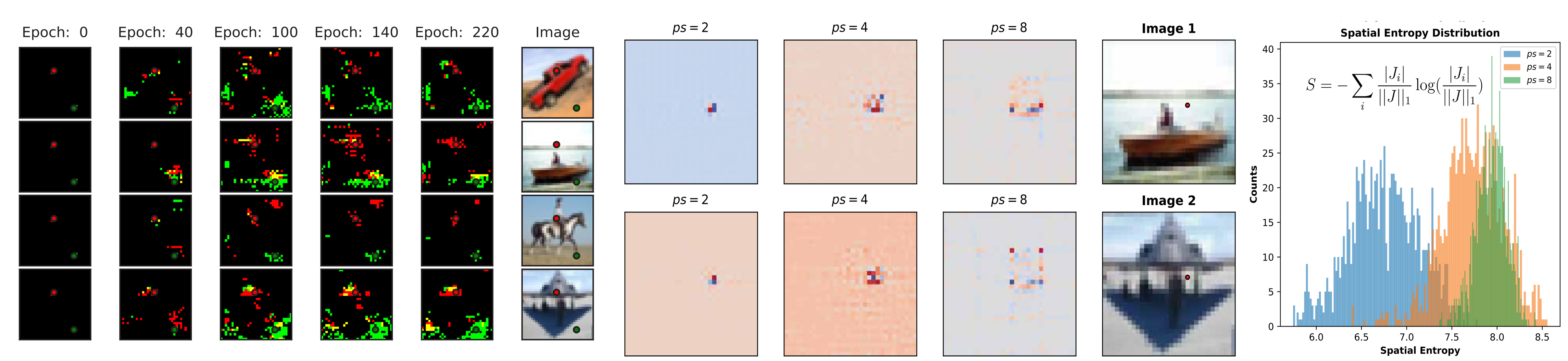
Key Insights:

- **Spatial Integration:** MAEs can integrate information from distant patches, whereas AEs and DAEs remain localized
- **Patch Size Controls Spatial Extent of the Reconstruction Kernel:** Larger patches lead to broader spatial receptive fields
- **Masking Ratio Controls Strength of the Regularizer:** Higher masking ratios lead to a stronger reliance on long-range spatial dependencies

MAEs capture spatial correlations in the data, with masking ratio and patch size controlling the spatial scale of the learned features



Characterizing the features of nonlinear MAEs



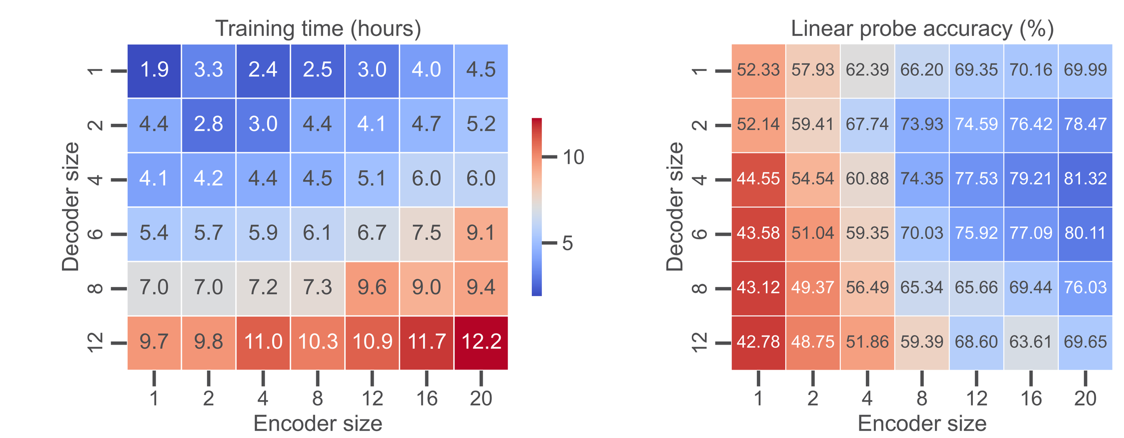
Jacobian magnitude averaged over inputs. Results shown for nonlinear models on CIFAR-10; similar behavior is observed for ImageNet

Key Insights:

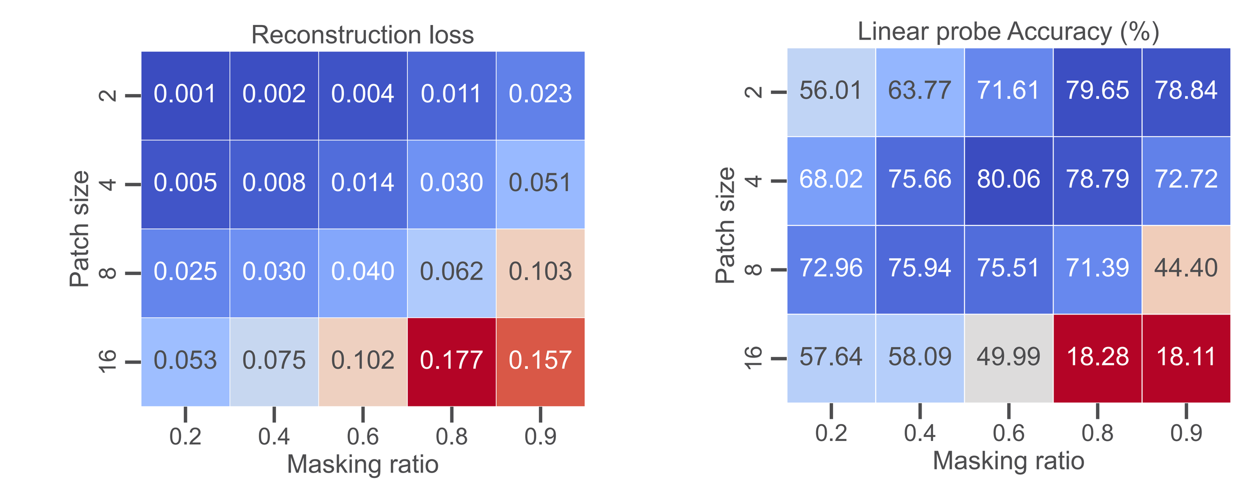
- **During Training:** Jacobians evolve from highly localized kernels to become spatially diffuse
- **Patch Size Controls the Spatial Extent of the Reconstruction Kernel:** Larger patch sizes yield reconstruction kernels with higher spatial entropy, shifting from local to global information aggregation
  - MAEs provide a potential mechanism for ViTs to learn local receptive fields

How to train your MAE

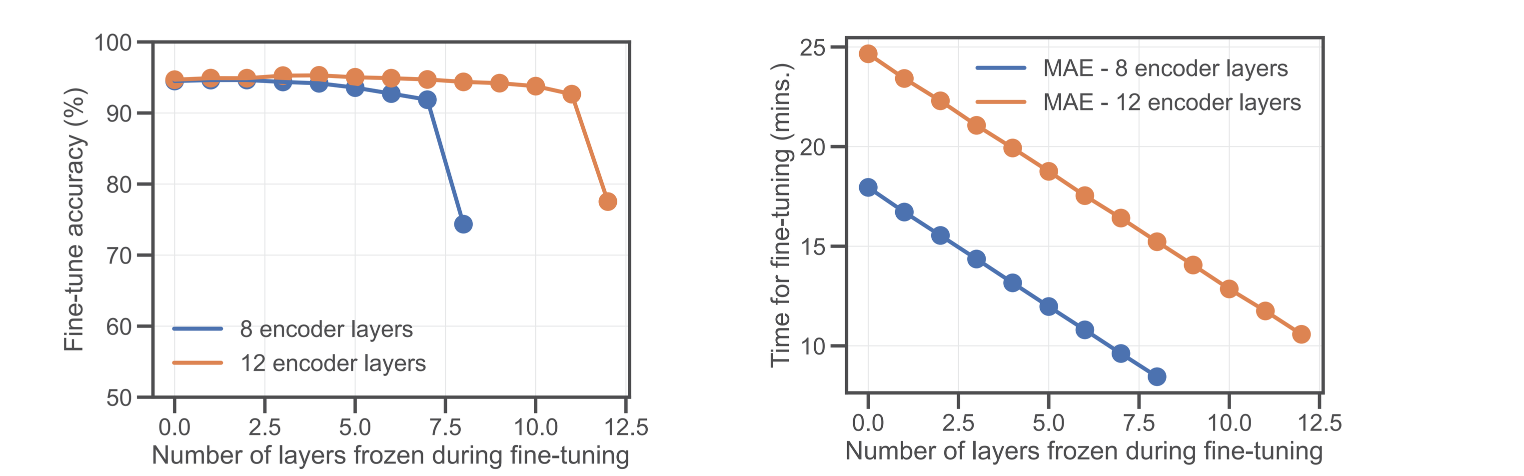
MAEs Benefit from Deeper Encoders and Minimal Decoders, Achieving Near Fine-Tuning Accuracy at 4× Faster Training Speed



Bigger Patches, Less Masking: Optimal MAE Performance Shifts Toward Lower Masking Ratios with Increasing Patch Size



Fine-Tuning Only a Few MAE Layers Achieves Near-Full Accuracy with Half the Training Cost



CIFAR-10 with 192 embedding dimensions pretrained for 2000 epochs with AdamW, and fine-tuned for 100 epochs

Conclusion

- **Hyperparameters determine the scale of the learned features:** Masking ratio and patch size set how broadly MAEs integrate spatial structure
- **Key Question**: How do spatial correlation scales relate to useful features for perception tasks?
  - For example, tasks such as optical flow require large spatial scales to overcome the aperture problem