Patch size and its effect on representations in MAEs

Anonymous Authors[1]

Abstract

Medical images contain a wealth of information that typically spans multiple spatial scales. While Masked Autoencoders (MAEs) are extensively used for self-supervised representation learning in these domains, they traditionally rely on a fixed patch size. In this work, we demonstrate that patch size inherently acts as a spatial filter, dictating the granularity of encoded features and causing models to struggle when representing objects smaller than the patch itself. To address this limitation and avoid the computational burden of training separate models for different resolutions, we introduce Mosaic MAE, an architecture trained dynamically across varying patch sizes and explicitly conditioned on patch scale via adaptive LayerNorm. Through experiments on a multi-scale MNIST-on-ImageNet dataset and lateral DXA bone scans, we demonstrate that Mosaic MAE offers improved representation learning relative to FlexiMAE—a self-supervised variant of FlexiViT. By decoupling feature learning from a single fixed resolution, Mosaic MAE enables the extraction of both fine-grained details and coarse semantic structures from a single pretrained model

1. Introduction

Standard MAEs (He et al., 2022) are trained with a fixed patch resolution (Feichtenhofer et al., 2022; Xie et al., 2022; Bao et al., 2021; Huang et al., 2022; Tong et al., 2022), implicitly assuming a uniform scale of interest. In domains such as medical imaging, this assumption falls short: critical information naturally spans multiple spatial scales, ranging from fine-grained textures (e.g., fused vertebrae) to coarse semantic landmarks (e.g., vertebral bodies). We hypothesize that the patch size p is a critical deterministic factor that governs the spatial scale of the features captured by the encoder. Specifically, the patch size acts similarly to a spatial filter, the network successfully encodes latent information about an object only when the patch size is smaller than the object itself. When vital features or objects are smaller than the designated patch size, the encoder struggles to encode them in its embeddings. The simplest alternative for capturing multi-scale representations requires training and deploying multiple separate models, each optimized for a specific patch size, which is computationally prohibitive and inefficient. To address this limitation, we introduce Mosaic MAE, a unified architecture trained dynamically across varying patch sizes. Drawing inspiration from multi-scale Vision Transformers (ViT) like FlexiViT (Beyer et al., 2023), we dynamically resize the linear projection weights and positional embeddings during training to accommodate randomly sampled patch sizes. Crucially, rather than forcing the model to implicitly learn a complex joint distribution of images and patch sizes $p(\boldsymbol{z}, p|\boldsymbol{x})$, we introduce explicit scale conditioning. By mapping the continuous patch size p to the scale and shift parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\beta}$) of the network's LayerNorm operations via an MLP (Perez et al., 2018; Peebles & Xie, 2023), Mosaic MAE learns to explicitly model the conditional distribution $p(\boldsymbol{z}|\boldsymbol{x}, p)$. This allows the model to actively recognize and adapt its internal representations to the current spatial scale. To validate our hypothesis and demonstrate the efficacy of Mosaic MAE, we present controlled experiments using a multi-scale benchmark of MNIST digits overlaid on ImageNet backgrounds, as well as real-world lateral vertebral assessment (LVA) scans from dual-energy X-ray absorptiometry (DXA) bone imaging.

Our approach builds upon several key advancements in representation learning. Mosaic MAE's dynamic sequence length is heavily inspired by architectures like FlexiViT (Beyer et al., 2023), which demonstrated that Vision Transformers can accommodate randomized patch sizes during supervised training via kernel interpolation. We extend this flexibility to the self-supervised generative regime. Finally, to ensure our model adapts to varying scales, we leverage conditional normalizations inspired by time-conditioning in DeiTs (Peebles & Xie, 2023) and adaptive LayerNorm (Perez et al., 2018) found in modern diffusion models (DiTs). This injects patch-size embeddings directly into the network, bridging the gap between dynamic

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

architecture sizing and scale-aware feature extraction.

In summary, our main contributions are as follows:

- We empirically demonstrate the "high-pass filter" effect of patch sizes on feature representations in standard MAEs, showing that fixed-patch models fail to encode objects smaller than their tokenization grid
- We introduce Mosaic MAE, a dynamic-patch architecture that leverages adaptive LayerNorm conditioning to explicitly model feature scale
- We show that explicit patch-size conditioning yields improved representations compared to unconditioned variable-patch baselines (FlexiMAE), successfully capturing multi-scale anatomical features in complex medical scans from a single pre-trained model

2. Patch size in MAEs

Let $\boldsymbol{x} \in \mathbb{R}^{h \times w \times c}$ denote an input image and let p denote the patch size. MAEs utilize an encoder-decoder Transformer architecture. An image is subdivided into $N_p = \lfloor h/p \rfloor \times \lfloor w/p \rfloor$ patches. Given a masking ratio m, only the unmasked patches are provided as input to the encoder. The masked tokens are then reintroduced prior to the decoder stage, and the network is trained to reconstruct the original pixel values of the masked patches using an ℓ_2 reconstruction loss. At inference time, only the encoder is retained to extract image embeddings, which is achieved either by averaging all patch-level tokens or by utilizing a dedicated $[cls]$ token. We hypothesize that the patch size p acts as a crucial hyperparameter that dictates the spatial scale of the features learned by the encoder. Concretely, the network's latent representations successfully encode object-level information better when the patch size is smaller than the object itself.

To validate this hypothesis, we construct a multi-scale MNIST-on-ImageNet dataset by superimposing random MNIST digits into one of the four corners of Tiny ImageNet images (Fig. 1). This setup allows for experimentation by varying two key parameters: the model patch size p and the spatial size s of the superimposed MNIST digit. We trained 9 distinct MAE models covering all combinations of $p \in \{4, 8, 16\}$ and $s \in \{8, 12, 14\}$, with the results presented in Fig. 1. We evaluate the representation quality using the linear probing accuracy for classifying the superimposed digits from the frozen encoder embeddings. The results clearly demonstrate that the linear probing accuracy is consistently higher when $p < s$ compared to when $p > s$, confirming that patch size fundamentally constrains the scale of encoded features. Consequently, when utilizing MAEs for representation learning, the patch size p functions as a predictable, "high-pass filter"-like mechanism to target features at specific scales. This property is particularly useful in domains such as medical and biological imaging, when the scales of target features (and those to be ignored) are known a priori.

3. Mosaic MAE

While one could train multiple independent models to determine the optimal patch size for a specific downstream task, this approach is computationally prohibitive. Instead, we propose training a single MAE with patch sizes uniformly sampled at random during the pre-training phase. This strategy ensures that, once trained, a single model can dynamically extract representations at varying patch resolutions during inference, depending on the specific scale requirements of the target task.

Accommodating variable patch sizes requires adapting the weights of the linear projection layer, which embeds the raw image patches into tokenized representations. Following the approach introduced in FlexiViT, we initialize these projection weights at a fixed spatial resolution of 32×32 and dynamically resize them to match the sampled patch size at each training step. Consistent with the observations in FlexiViT, we found that the specific dimension of this initial weight matrix does not significantly impact overall performance. Furthermore, varying the patch size inherently alters the sequence length N_p of the tokenized image. To handle this, we initialize a fixed-size positional embedding and dynamically resize it to the target sequence length using bi-linear interpolation.

Previous literature demonstrates that generative models often optimize more effectively when modeling conditional distributions rather than complex joint distributions. We hypothesize that it is easier for the network to learn the distribution of latent representations conditioned explicitly on the patch size, $p(\boldsymbol{z}|\boldsymbol{x}, p)$, rather than implicitly modeling the joint distribution $p(\boldsymbol{z}, p|\boldsymbol{x})$. To achieve this, we introduce explicit patch-size conditioning during MAE pre-training. Drawing inspiration from time-step conditioning in diffusion models and DeiTs, the continuous patch size p_i is mapped via a Multi-Layer Perceptron (MLP) to the scale ($\boldsymbol{\beta}$) and shift ($\boldsymbol{\mu}$) parameters of the LayerNorm operations across all encoder and decoder layers l:

$$\boldsymbol{\mu}_i^l, \boldsymbol{\beta}_i^l = f_{\text{MLP}}(p_i)$$
$$\boldsymbol{z}_i^l = \left(\frac{\boldsymbol{z}_i^l - \mathbb{E}[\boldsymbol{z}_i^l]}{\sigma(\boldsymbol{z}_i^l)} \right) \odot \boldsymbol{\beta}_i^l + \boldsymbol{\mu}_i^l$$

At inference time, representations are extracted by simply providing the image alongside the desired patch size and its corresponding conditioning signal. However, because the encoder must capture this conditioning information to produce scale-aware features and the decoder is ultimately discarded during representation retrieval we hypothesize

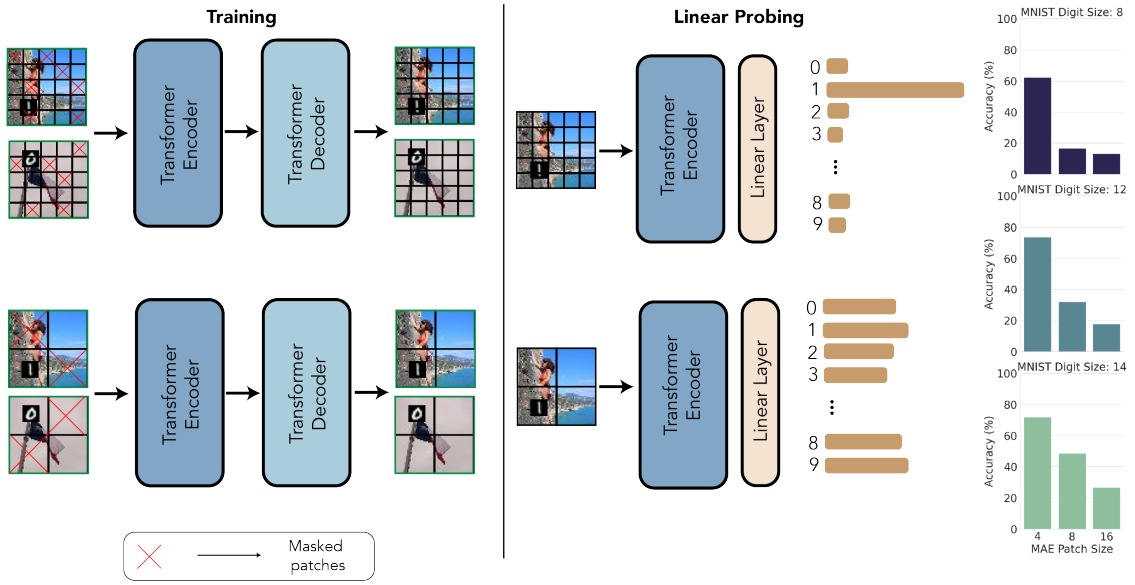

Figure 1. Left & Middle: MAE pre-training on the MNIST-on-ImageNet benchmark. Smaller patches (top) preserve fine-grained spatial information. Conversely, larger patches (bottom) spatially integrate over small targets, resulting in diffuse representations. Right: Downstream linear probing accuracy across varying MNIST digit sizes ($s \in \{8, 12, 14\}$) and patch sizes ($p \in \{4, 8, 16\}$). Performance consistently degrades as the MAE patch size increases, demonstrating that encoders struggle to resolve objects smaller than their tokenization grid

that applying patch conditioning to the decoder may inadvertently degrade the encoder's representation quality. We evaluate this hypothesis in our subsequent experiments.

4. Experiments

We evaluate our proposed Mosaic MAE through experiments investigating three questions: (1) How does patch size fundamentally affect multi-scale feature learning? (2) Does explicit patch-size conditioning yield better downstream representations? (3) Does decoder patch size conditioning degrade Mosaic MAE's representations?

4.1. Setup

We utilize Tiny ImageNet (Le & Yang, 2015) (200 classes, 64×64 resolution) augmented with MNIST digits (LeCun et al., 2010) overlaid at three distinct spatial scales: 6×6, 8×8, and 14×14 pixels. This constructs a multi-scale benchmark requiring models to simultaneously capture fine-grained digit features and coarse background textures. To prevent localization biases, overlay coordinates are fixed via pre-computed mappings.

We compare four primary model families: (1) **Base MAE** (He et al., 2022) trained with fixed patch sizes $p \in \{4, 8, 16\}$, (2) **FlexiMAE** (Beyer et al., 2023), an unconditioned baseline trained with variable patch sizes $\{4, 8, 16\}$ utilizing kernel interpolation, (3) **Mosaic MAE** (ours), which explicitly conditions on patch size via sinusoidal embeddings injected through adaptive LayerNorm (Peebles & Xie, 2023), and (4) **Mosaic MAE w/ Dec Cond**, which extends (3) by applying scale conditioning to both the encoder and the decoder. All models utilize a ViT-Small backbone (embedding dimension 384). Pre-training is conducted for 100 epochs with a mask ratio of 0.75, utilizing the AdamW optimizer (lr=10^{-3}, wd=0.05), a batch size of 256, and a cosine learning rate schedule with a 40-epoch warmup, optimizing a normalized pixel reconstruction loss.

4.2. Analysis

After self-supervised pre-training, we evaluate representation quality via linear probing on a 10-way MNIST digit classification task. This metric isolates whether the frozen learned representations successfully capture features at specific target scales. For variable-patch models, we report inference accuracy across all three patch sizes $\{4, 8, 16\}$. Table 1 details the MNIST digit classification accuracy across the three overlay scales.

As expected, fixed-scale Base MAE models perform optimally only when the evaluation patch size aligns closely with the target feature scale. For instance, Base MAE-P4 excels on fine-grained (6×6) targets, while Base MAE-P16 fails at capturing fine details. Training with variable patch sizes (FlexiMAE) yields performance close to fixed patch size models, but there is a significant gap between them.

Mosaic MAE on average outperforms the unconditioned variable-patch baseline. By explicitly modeling scale via

Table 1. MNIST digit classification accuracy across multiple overlay sizes. All models use ViT-S (dim 384) pretrained for 100 epochs. Linear probing evaluates 10-way digit classification. Variable-patch models are evaluated at three inference patch sizes

Model	Patch Size	MNIST 6×6	MNIST 8×8	MNIST 14×14	Mean
Base MAE-P4	4	39.45	59.55	71.97	56.99
Base MAE-P8	8	11.80	16.44	48.65	25.63
Base MAE-P16	16	11.96	12.26	26.65	16.96
Flexi-MAE	4	19.83	41.42	60.82	40.69
Flexi-MAE	8	14.21	16.13	48.39	26.24
Flexi-MAE	16	12.63	12.25	28.98	17.95
Mosaic MAE w/ Dec Cond	4	21.93	34.69	55.17	37.26
Mosaic MAE w/ Dec Cond	8	14.11	16.36	47.79	26.08
Mosaic MAE w/ Dec Cond	16	12.27	11.92	22.65	15.61
Mosaic MAE	4	34.12	39.28	70.80	48.07
Mosaic MAE	8	16.83	25.82	61.34	34.66
Mosaic MAE	16	12.12	11.96	18.69	14.26

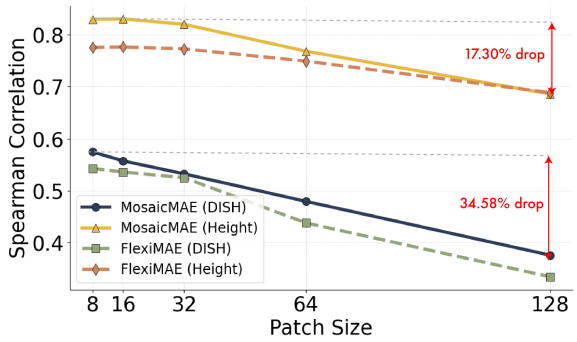

Figure 2. Downstream linear probing performance on clinical DXA scans. We report the Spearman correlation for predicting patient Height and DISH scores across a wide range of inference patch sizes (8 to 128)

learnable LayerNorm parameters, the encoder dynamically adjusts its feature extraction processes, yielding the best average accuracy and better cross-scale generalization. Contrary to the encoder-only variant, injecting scale information into the decoder (Mosaic MAE w/ Dec Cond) markedly degrades downstream linear probing performance. This validates our earlier hypothesis: when the decoder shares the burden of scale-awareness, the encoder is disincentivized from capturing robust, standalone scale-invariant representations in its latent space.

4.3. DXA scans

To demonstrate generalization beyond natural image benchmarks, we evaluate Mosaic MAE on lateral vertebral assessment (LVA) scans of DXA from UKBB. Images are preprocessed via cropping to the vertebral column region and resized to 256×256 pixels. Standard augmentations (rotation $\pm 10°$, 80% random crops, color jittering) are applied to improve robustness. Models are pre-trained for 150 epochs with a batch size of 32, dynamically sampling patch sizes $p \in \{8, 16, 32, 64, 128\}$ to match the multi-scale nature of the skeletal structures.

Rather than relying solely on reconstruction loss, we assess representation quality by predicting two quantities that requires different scale of features to be learnt, specifically patient Height and Diffuse Idiopathic Skeletal Hyperostosis (DISH) score (Sethi et al., 2023), via linear probing on the frozen encoder embeddings. Fig. 2 illustrates the Spearman correlation of these predictions across an expanded range of inference patch sizes ($p \in \{8, 16, 32, 64, 128\}$).While both models naturally experience a degradation in predictive power at extreme patch resolutions, Mosaic MAE demonstrates better scale robustness compared to the unconditioned FlexiMAE baseline for both Height and DISH score. In addition, when looking at the degradataion of performance between the largest ($p = 128$) and smallest ($p = 8$) patch sizes, fine grained DISH score prediction drops roughly $2\times$ as much as a coarse grained feature like height. This indicates that larger patch sizes suffice for encoding coarse-grained features, whereas smaller patch sizes are necessary when fine-grained features are important. Mosaic MAE offers this flexibility from a single training run.

5. Conclusion

By explicitly conditioning variable-patch training via adaptive LayerNorm, Mosaic MAE overcomes some of the spatial filtering limitations of fixed-resolution Masked Autoencoders, enabling the robust extraction of both fine-grained and coarse multi-scale representations from a single pretrained model.

References

Bao, H., Dong, L., Piao, S., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Beyer, L., Izmailov, P., Kolesnikov, A., Caron, M., Kornblith, S., Zhai, X., Minderer, M., Tschannen, M., Alabdulmohsin, I., and Pavetic, F. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14496–14506, 2023.

Feichtenhofer, C., Li, Y., He, K., et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and 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.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen. *Advances in neural information processing systems*, 35:28708–28720, 2022.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Sethi, A., Ruby, J. G., Veras, M. A., Telis, N., and Melamud, E. Genetics implicates overactive osteogenesis in the development of diffuse idiopathic skeletal hyperostosis. *Nature Communications*, 14(1):2644, 2023.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and 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.