From Pixels to Components: Eigenvector Masking for Visual Representation Learning

Alice Bizeul[1,2] **Thomas M. Sutter**[1] **Alain Ryser**[1] **Bernhard Schölkopf**[2] **Julius von Kügelgen**[3] **Julia E. Vogt**[1]

Abstract

Predicting masked from visible parts of an image is a powerful self-supervised approach for visual representation learning. However, the common practice of masking random patches of pixels exhibits certain failure modes, which can prevent learning meaningful high-level features, as required for downstream tasks. We propose an alternative masking strategy that operates on a suitable transformation of the data rather than on the raw pixels. Specifically, we perform principal component analysis and then randomly mask a subset of components, which accounts for a fixed ratio of the data variance. The learning task then amounts to reconstructing the masked components from the visible ones. Compared to local patches of pixels, the principal components of images carry more global information. We thus posit that predicting masked from visible components involves more high-level features, allowing our masking strategy to extract more useful representations. This is corroborated by our empirical findings which demonstrate improved image classification performance for component over pixel masking. Our method thus constitutes a simple and robust data-driven alternative to traditional masked image modeling approaches[†].

1. Introduction

In masked image modeling (MIM; Pathak et al., 2016), parts of an image are masked, and a model has to reconstruct the missing parts from the visible ones—analogous to predicting hidden words in masked language modeling (Devlin, 2019). To succeed at this task, it is thought that the model must learn a meaningful representation of the visual content (Kong et al., 2023). Empirically, this approach indeed produces representations that perform well when fine-tuned on downstream tasks, such as image classification and semantic segmentation (Zhou et al., 2021; Bao et al., 2022; Xie et al., 2022; Baevski et al., 2022; Dong et al., 2023).

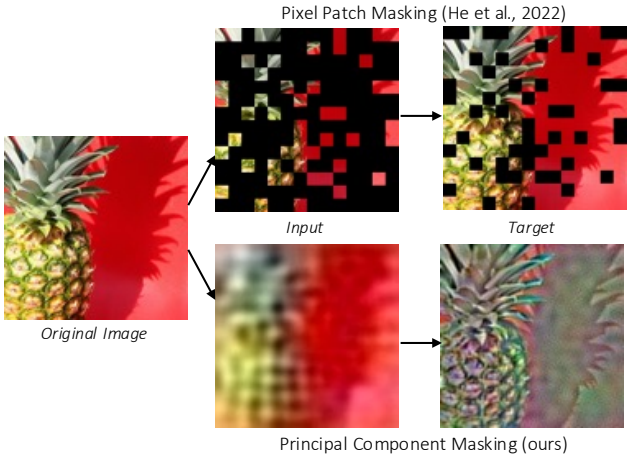

Figure 1. **From Pixels to Components.** Masked image modeling involves reconstructing masked-out patches of pixels from visible ones. Instead of masking in pixel space (top), we propose applying random masks to a transformed version of the image, specifically to its principal component representation (bottom). Two disjoint sets of components are used as input and reconstruction target.

MIM has been particularly effective when combined with vision transformers (ViT; Dosovitskiy et al., 2021). A prominent example is the masked autoencoder (MAE; He et al., 2021), consisting of a ViT encoder-decoder architecture and a masking strategy that randomly selects a fixed ratio of square image patches, see Figure 1 (top). The encoder processes the visible patches and the decoder aims to reconstruct the masked content from the inferred representation.

While the inner workings of MAEs remain poorly understood (Zhang et al., 2022; Yue et al., 2023), Kong et al. (2023) suggested an explanation from a latent variable model perspective. By splitting an image into two separate views and asking the model to predict one from the other, MAEs are compelled to pick up any information that is shared across views. If the partition into views (i.e., the masking strategy) is chosen *carefully*, this shared information will include high-level latent variables, such as object class. Since solving common downstream tasks with a simple (e.g., linear) predictor precisely requires identifying such

[†]We release the code to reproduce our results on GitHub. Correspondence to alice.bizeul@inf.ethz.ch
[1] Department of Computer Science, ETH Zürich [2] MPI for Intelligent Systems, Tübingen [3] Seminar for Statistics, ETH Zürich

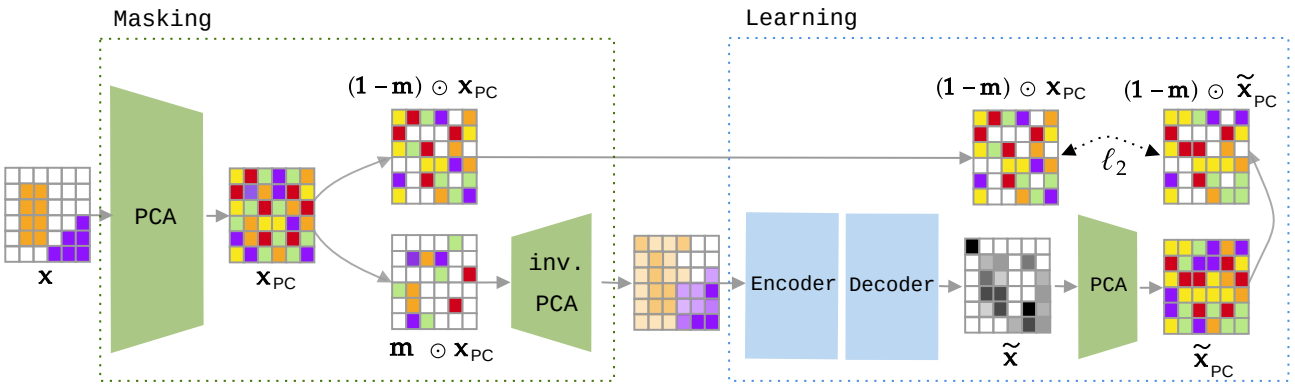

Figure 2. **Overview of the Principal Masked Autoencoder.** A principal masked autoencoder (PMAE) differs from a vanilla MAE (He et al., 2021) by performing the masking in the space of principal components $\mathbf{x}_{\text{PC}} = \text{PCA}(\mathbf{x})$ rather than in pixel space. The visible principal components $\mathbf{m} \odot \mathbf{x}_{\text{PC}}$ are then projected back into the observation space and serve as the input for a ViT encoder-decoder architecture. Masked principal components, $(\mathbf{1}-\mathbf{m}) \odot \mathbf{x}_{\text{PC}}$, serve as the reconstruction target.

high-level features, this offers a possible explanation for the observed effectiveness of MAE representations.

With this in mind, we ask: *Is masking random patches in pixel space the optimal strategy for MIM?* In natural language, each word in a sentence tends to carry semantic information, and the information shared between sets of words often conveys the general message of the sentence. However, this does not necessarily apply to the visual domain, where individual pixels or entire patches commonly contain redundant information (e.g., background). Moreover, (small) objects can be masked out completely such that any information about them is lost and reconstruction becomes impossible, see Figure 3 (right). Some works have thus sought to devise more elaborate masking strategies which leverage auxiliary information such as learned or given image segmentations (Li et al., 2021; Kakogeorgiou et al., 2022; Shi et al., 2022). Without such prior knowledge or complex training pipelines to identify the structure of an image, randomly masking a fixed ratio of patches of pixels remains the default practice. Yet, relying on this masking strategy assumes—rather unrealistically—that the information shared between any random partition of pixel patches naturally aligns with high-level variables of interest (Kong et al., 2023).

In this work, we introduce an *alternative data-driven masking strategy for MIM*. Rather than working directly in pixel space, we propose to first project images into a latent space and then perform random masking on the transformed data. Among the infinitely many candidate transformations, we opt for simplicity and choose principal component analysis (PCA), a well-established linear pre-processing technique that is deterministic and (hyper)parameter-free. Following the PCA transformation, some principal components are masked-out and need to be reconstructed from the remaining visible components, see Figure 1 (bottom) for an example. Further, we leverage the fact that each principal component captures a known fraction of the data variance to inform our masking strategy. Specifically, we mask a random subset of principal components that accounts for a fixed ratio of the variance (Figure 4). The ratio of masked variance serves as a proxy for the complexity of the modeling task making it a more interpretable and more easily tunable hyperparameter than the ratio of masked patches (Figure 5, left). We combine this masking strategy with the MAE architecture and refer to the resulting method as *principal masked autoencoder* (PMAE), see Figure 2 for an overview.

Relying on a transformation like PCA allows for partitioning the information in an image into a set of global features rather than into local patches of pixels, see Figure 1 . This can help overcome the aforementioned failure modes of spatial masking where the input and target share too much (redundancy) or too little (impossibility) information. The space of principal components may, therefore, constitute a more meaningful domain for masking, resulting in more useful high-level representations.

This is consistent with recent work, highlighting the beneficial partitioning of image information by PCA. Balestriero & LeCun (2024) demonstrate that low-eigenvalue components capture features crucial for common downstream tasks (see also Figure 7), and Chen et al. (2024b) highlight the importance of the space in which image distortions are applied, referring to PCA as a valuable transformation to consider. To the best of our knowledge, our work is the first to leverage such insights to devise a simple, robust, and effective data-driven alternative to pixel-space masking in MIM.

In experiments on the CIFAR10, TinyImageNet and three medical MedMNIST datasets, we observe our approach (PMAE) to yield superior performance to the default strategy of spatial masking (MAE), while being less sensitive to the choice of masking ratio hyperparameter. These empirical findings support our claim that masking principal components instead of pixel patches can facilitate the learning of more meaningful high-level representations.

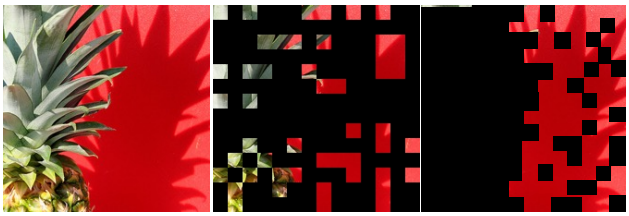

Figure 3. **Masking in pixel space.** Image (left) with a random spatial mask *partially* removing relevant information (middle) and removing *all* semantic information (right). The latter is an example in which MIM would likely fail to learn useful representations.

2. Background

Principal Component Analysis.

Principal component analysis (PCA; Pearson, 1901; Hotelling, 1933) identifies components in data that exhibit the highest variance. Given a centered data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$ with N observations of dimension D, PCA seeks weight vectors $\mathbf{v}_l \in \mathbb{R}^D$ for $l = 1, \ldots, L \leq D$ (the principal components or PCs) that maximize the variance of the projections $\mathbf{X}\mathbf{v}_l$, subject to orthogonality with previously found PCs and unit-length constraints. The solution to this problem is given by the eigenvalue decomposition of the empirical covariance matrix $\mathbf{\Sigma} = \mathbf{X}^\top \mathbf{X}$, i.e., $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_D)$ contains the eigenvalues $\lambda_1 > \ldots > \lambda_D$, and $\mathbf{V} \in \mathbb{R}^{D \times D}$ contains the corresponding eigenvectors. The first L PCs correspond to the eigenvectors of $\mathbf{\Sigma}$ with the largest eigenvalues, capturing the dominant modes of variation in the data, with the variance explained by each PC proportional to its eigenvalue λ_l. While PCA is often used with $L < D$ for dimensionality reduction, we focus on the lossless case where $L = D$. The projection of $\mathbf{X}$ onto its principal components ("into PC space") is given by $\mathbf{X}_{\text{PC}} = \mathbf{X}\mathbf{V}$, and the inverse transformation by $\mathbf{X} = \mathbf{X}_{\text{PC}}\mathbf{V}^\top$. By construction, PCA produces features which are uncorrelated ($\mathbf{X}_{\text{PC}}^\top \mathbf{X}_{\text{PC}} = \mathbf{\Lambda}$). However, we emphasize that—except for special cases such as when the data is multivariate Gaussian—this does *not* imply independence. In general, the values of a subset of PCs are therefore (non-linearly) predictive of the values of other PCs. Our results in Section 5 and examples of reconstructions of masked principal components in Appendix A.2.1 demonstrate that this is the case for the natural and medical images we consider.

Representation Learning. Representation learning (Bengio et al., 2013) aims at learning an embedding function or encoder $f : \mathbf{x} \mapsto \mathbf{z}$, which maps observations $\mathbf{x} \in \mathbb{R}^D$ to representations $\mathbf{z} \in \mathbb{R}^K$. These representations are meant to capture some of the explanatory factors underlying the data, thus making them well-suited for use in downstream tasks such as predicting a target variable y (e.g., the class or location of objects), often thought of as simple functions of the high-level explanatory factors.

Masked Image Modeling. Prominent approaches to representation learning in the image domain rely on the masked image modeling paradigm (Pathak et al., 2016; Zhou et al., 2021; He et al., 2021; Bao et al., 2022). Here we focus on the widely adopted MAE (He et al., 2021) as a representative of MIM. A MAE consists of: an encoder $f_{\boldsymbol{\phi}}$, parametrized by $\boldsymbol{\phi}$, which maps the visible portions of the input together with their positional embeddings to a representation; and a decoder $g_{\boldsymbol{\theta}}$, parametrized by $\boldsymbol{\theta}$, which reconstructs the missing parts from their positional embeddings and the representation produced by the encoder.

Given an observation $\mathbf{x} \in \mathbb{R}^D$ and complementary binary masks $\mathbf{m}, (\mathbf{1} - \mathbf{m}) \in \{0, 1\}^D$, which extract the visible and masked parts, respectively, the MAE objective is given by

$$\mathcal{L}_{\text{MAE}}(\mathbf{x}, \mathbf{m}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \left\| (\mathbf{1} - \mathbf{m}) \odot \left[g_{\boldsymbol{\theta}} \circ f_{\boldsymbol{\phi}} \left(\mathbf{m} \odot \mathbf{x} \right) - \mathbf{x} \right] \right\|_2^2, \tag{2.1}$$

where $\odot$ denotes element-wise multiplication and $\circ$ denotes function composition. The parameters $(\boldsymbol{\phi}, \boldsymbol{\theta})$ are optimized via stochastic gradient descent on Equation (2.1).

The mask $\mathbf{m}$ partitions the D pixels into two disjoint sets of $(1 - r)D$ visible and rD masked out pixels, where r is referred to as the *masking ratio*. Typically, $\mathbf{m}$ is chosen by partitioning the image into square patches (thus introducing patch size as an additional hyperparameter) and then randomly selecting a ratio of r patches to be masked.

Prior work has relied on hyperparameter sweeps to identify the masking ratio and patch size that optimize downstream performance (He et al., 2021; Zhang et al., 2022). These efforts have led to the widely adopted approach of masking out 75% ($r = 0.75$) of patches of size 16×16 pixels.

Challenges Resulting from Spatial Masking. Even though MIM with random spatial masking produces strong results on representation learning benchmarks (Dong et al., 2023), it is based on the rather unrealistic assumption that, given any partition of image patches into two disjoint sets, the information shared between views contains variables that are linearly predictive of downstream targets y (Kong et al., 2023). As visualized in Figure 3, for some masks (such as the one in the middle) shared information indeed includes features such as object type. However, for other masks (such as the one on the right) predicting the correct class label from the visible patches is almost impossible. The latter therefore yields input-target pairs which do not contribute a useful self-supervised learning signal. Moreover, even for well-designed masks, many masked-out and visible patches contain redundant information such as background. This leads us to conjecture that spatial masking is a suboptimal approach to MIM that is not always well-aligned with common downstream tasks and may suffer from slow convergence and high sensitivity to hyperparameters, as

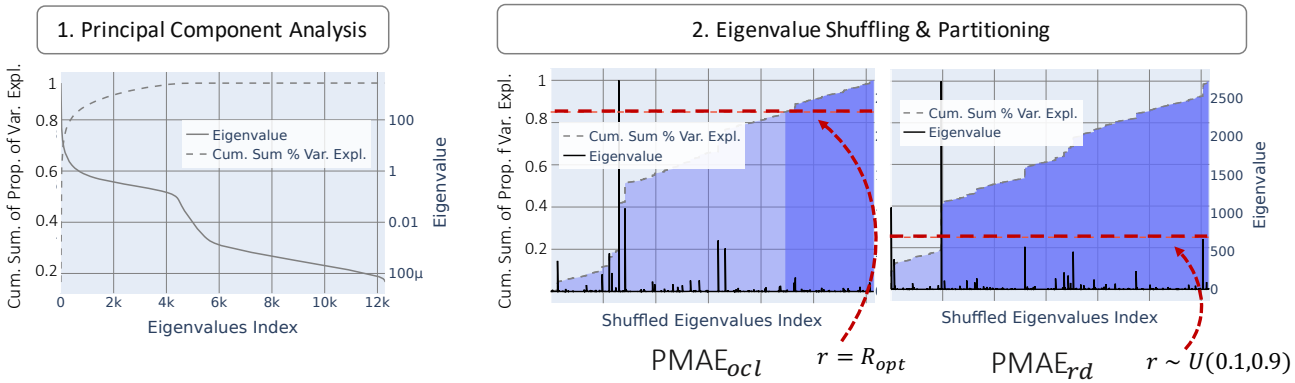

Figure 4. **Mask Design in PMAE.** 1. Perform PCA 2. For each batch, randomly shuffle the principal components and select a subset to construct the input (light blue), while the remaining components are used to create the reconstruction target (dark blue). In PMAE$_{\text{ocl}}$, the input components are chosen to explain $100\times(1-r)\%$ of the data's variance with r optimized for downstream performance (here $R_{\text{opt}} = 0.15$). In PMAE$_{\text{rd}}$, the input explains between 10% and 90% of the variance with r sampled independently and uniformly in $[0.1, 0.9]$ for each batch.

suggested by prior work (He et al., 2021; Balestriero & LeCun, 2024) and confirmed in Figure 5.

3. Principal Masked Autoencoders

To address the challenges arising from spatial masking, we propose an alternative masking strategy and a corresponding MAE-variant which we refer to as *principal masked autoencoder* (PMAE). PMAE builds on the MIM learning paradigm but differs from prior approaches by performing the masking operation on a latent space. Specifically, we first apply an invertible transformation $t : \mathbb{R}^D \to \mathbb{R}^D$ to an image $\mathbf{x}$ and then mask and reconstruct at the level of $t(\mathbf{x})$, resulting in the following objective:

$$\mathcal{L}_{\text{PMAE}}(\mathbf{x}, \mathbf{m}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \left\| (\mathbf{1} - \mathbf{m}) \odot \left[t \circ g_{\boldsymbol{\theta}} \circ f_{\boldsymbol{\phi}} \circ t^{-1}(\mathbf{m} \odot t(\mathbf{x})) - t(\mathbf{x}) \right] \right\|_2^2 .$$

Similar to Equation (2.1), $f_{\boldsymbol{\phi}}$ is an embedding function which encodes visible parts of the image and $g_{\boldsymbol{\theta}}$ is a decoder which reconstructs the missing parts. Note that Equation (2.1) is recovered as a special case of Equation (3.1) if t is chosen to be the identity mapping.

Design Choices. While Equation (3.1) can generally accommodate any invertible mapping t, this work specifically explores the use of PCA as a data transformation, i.e., we choose $t(\mathbf{x}) = \mathbf{x}_{\text{PC}} = \mathbf{x}\mathbf{V}$. Figure 2 provides a visual summary of PMAE based on PCA. We leave an investigation of alternative image transformations (e.g., the Fourier transform) for future work, see Section 8 for further discussion.

Following MAEs, we use ViTs (Dosovitskiy et al., 2021) for both $f_{\boldsymbol{\phi}}$ and $g_{\boldsymbol{\theta}}$. Since these architectures are optimized for processing images, we project the visible components back into image space using the inverse PCA transformation t^{-1} before encoding. Similarly, to reconstruct the missing components, the decoder output is projected back into PC space using t. ViTs are a common choice of architecture for MIM due to the use of binary masks in pixel space. These masks produce images with black-patch occlusions (see Figure 3), which are poorly aligned with the inductive biases of other architectures. By masking principal components instead of patches of pixels, the encoder inputs and reconstruction targets in PMAEs display smoother activation patterns (see Figure 1, bottom). PMAEs are therefore, in principle, also compatible with, e.g., convolutional networks. An exploration of other architectural choices for PMAE is left for future work.

Analogous to the MAE objective in Equation (2.1), the PMAE objective only penalizes reconstruction errors on the part masked out by $(\mathbf{1} - \mathbf{m})$. In Equation (3.1), the model output is projected into PC space, and the l_2 norm is computed between the predicted and ground truth masked-out principal components. In Appendix A.2.4, we explore an alternative objective for which the reconstruction error is calculated directly in the image space. Specifically, Equation (A.1) computes the l_2 norm between the decoder output and the masked-out principal components projected back into the image space. However, this alternative approach results in lower downstream performance, likely due to the stricter constraints placed on the decoder which, in this case, also needs to predict the effect of the visible components—akin to penalizing an MAE decoder for in-painting, rather than predicting perfectly black patches for the visible parts.

Mask design is also a crucial aspect of our approach. In MAEs, a fixed percentage of pixel patches is masked out. The masking ratio is usually chosen based on a hyperparameter tuning conducted for each individual dataset (see Figure 5, top). Instead in PMAE, we choose to systematically mask components that collectively account for a fixed percentage of the variance of the dataset. Figure 4 illustrates this process. After applying PCA, we randomly shuffle the order of principal components within each batch and divide

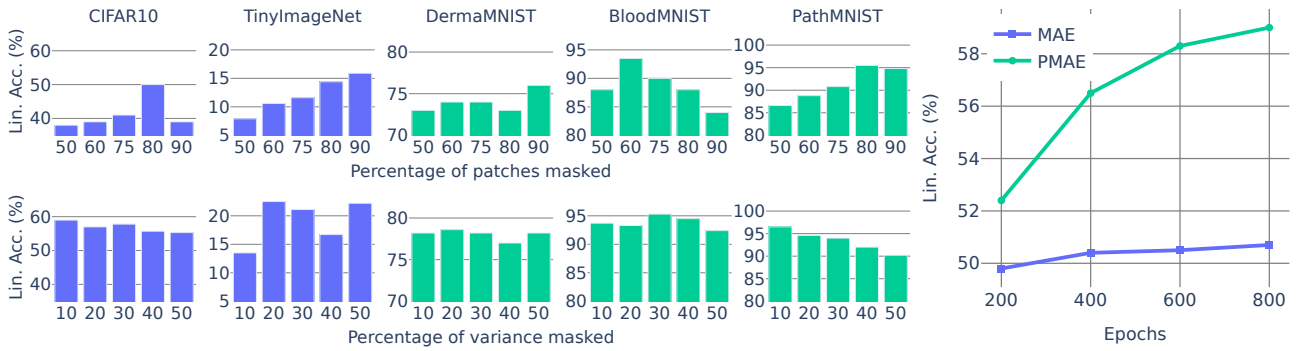

Figure 5. **Impact of the Masking Ratio.** MAE (top) and PMAE (bottom) linear probing accuracy for varying masking ratios. The masking ratio is a sensitive and data-dependent hyper-parameter. While for MAE a clear masking guideline is hard to extract, for PMAE we observe close-to-optimal performance across datasets for 20% of the data variance masked. **(Right) Learning curves.** Linear probe accuracy for CIFAR10 classification across training epochs. PMAE outperforms MAE's final performance after 200 epochs.

them into two disjoint sets: one explaining $100(1-r)\%$ of the variance, and the other, $100r\%$, where r is our masking ratio. The ratio of masked variance serves as a more interpretable and easily tunable hyperparameter compared to the masked patch ratio used in MAEs (Figure 5, bottom), effectively acting as a proxy for modeling task complexity. Further details on the choice of r are provided in Section 4.

Intuition behind PMAE. Contrary to spatial masking presented in Section 2, an appropriate function t, in Equation (3.1), can encourage information that is shared between visible and masked-out information to contain y. More specifically, if the latent space captures unique global information in each dimension, masking any of these dimensions retains information about all parts of the image. Hence, Equation (3.1) allows us to learn more meaningful representations for a suitable choice of t. While there may be many appropriate choices for t, we found that applying PCA and projecting samples using the resulting principal components is a suitable choice for the latent space. In particular, each dimension captures specific factors of variation observed within the dataset and is typically tied to global features as shown in Figure 7. Masking one factor of variation thus prevents us from completely removing all information about variables of interest within a sample, as most principal components will retain some information about them.

In PMAE, the masking ratio—the proportion of explained variance masked—can also easily be interpreted, as it directly corresponds to a well-understood quantity. In contrast, masking pixel patches lacks a straightforward connection to a well-defined metric. Depending on the dataset, a given proportion of masked patches can correspond to varying amounts of information content, making the masking ratio in the spatial domain harder to interpret.

4. Experimental Setup

We now outline the setup used to validate PMAE. Our experiments adapt the implementation proposed by He et al. (2021). Further experimental details and information about computational costs can be found in Appendix A.1.

Mask Design. Following MAE (He et al., 2021), we take random image patch masking as our baseline. Based on ablation studies from He et al. (2021), the standard practice involves masking out 75% of image patches (denoted as MAE_{std}). We also examine an oracle-based masking strategy (denoted as MAE_{ocl}), where the masking ratio is tuned to optimize linear probing downstream performance. This setting serves as an upper bound to the downstream performance. Additionally, we introduce a randomized masking approach, MAE_{rd}, in which the masking ratio is independently sampled within a range of 0.1 to 0.9 for each batch. This strategy is exempt from any hyperparameter tuning and offers insights into the robustness of methods under suboptimal hyperparameters.

A similar approach is applied to PMAE, where we consider both oracle (PMAE_{ocl}) and randomized (PMAE_{rd}) masking strategies. In the oracle approach, we define the optimal percentage of *variance* to be masked based on linear probe downstream performance on a held-out dataset. In the randomized strategy, we ensure at least 10% and at most 90% of the data variance is masked out. This percentage is independently sampled for each batch. Figure 9 provides examples of the images obtained from these masking strategies.

Training & Evaluation. We train a ViT-Tiny (Touvron et al., 2021; Dosovitskiy et al., 2021) encoder and decoder backbones. Following He et al. (2021), we use image flipping and random image cropping as data augmentations. The decoder's output is normalized (He et al., 2021) for MAEs only as this did not result in a consistent performance gain for PMAE. We train representations for 800 epochs

Table 1. Linear & MLP probe and fine-tuning top-1% accuracy for CIFAR10, TinyImageNet and MedMNIST datasets for random masking in pixel (MAE) and principal component (PMAE) space with the standard 75% (std), oracle (ocl) and random (rd) masking ratios. * refers to ours.

		CIFAR10	TinyIN	Derma	Blood	Path
Linear	MAE_{std}	41.7	11.5	72.4	73.4	83.4
	MAE_{ocl}	50.7	15.5	73.7	78.6	86.4
	$\text{PMAE}^\star_{\text{ocl}}$	**59.0**	**22.5**	**78.6**	**95.5**	**96.8**
	MAE_{rd}	41.9	7.5	72.4	83.2	85.6
	$\text{PMAE}^\star_{\text{rd}}$	**44.0**	**16.9**	**76.4**	**90.0**	**90.1**
MLP	MAE_{std}	34.0	15.5	72.2	68.6	92.6
	MAE_{ocl}	55.2	22.2	74.4	75.8	95.1
	$\text{PMAE}^\star_{\text{ocl}}$	**64.1**	**25.1**	**80.2**	**92.5**	**98.6**
	MAE_{rd}	38.5	11.6	66.9	70.6	95.7
	$\text{PMAE}^\star_{\text{rd}}$	**47.0**	**22.6**	**77.4**	**80.2**	**97.7**
Fine-tuned	MAE_{std}	75.7	37.5	80.4	97.8	**99.7**
	MAE_{ocl}	80.5	42.8	79.9	**98.1**	**99.7**
	$\text{PMAE}^\star_{\text{ocl}}$	**84.8**	**44.5**	**82.3**	**98.1**	**99.7**
	MAE_{rd}	77.3	39.7	79.6	97.4	99.6
	$\text{PMAE}^\star_{\text{rd}}$	**80.4**	**46.9**	**82.4**	**98.5**	**99.7**

and provide an overview of the evolution of performance across training in Appendix A.2. We then evaluate learned representations on image classification using a linear probe and multi-layer perceptron (MLP) classifier on top of the encoder's output `[CLS]` token which is frozen. Additionally, we also explore the fine-tuning setting in which the pre-trained encoder and a linear probe appended to the `[CLS]` token are trained simultaneously. Following He et al. (2021), we fix the training duration at evaluation to 100 epochs. Appendix A.2 also reports downstream performance obtained with a k-NN classifier.

5. Results

In this section, we will outline and analyze the empirical advantages of PMAE compared to standard MAEs in image classification tasks. Specifically, we provide evidence that masking within the space of principal components facilitates the learning of discriminative features, resulting in improved classification performance. Our findings are supported by empirical evidence across five datasets, including two natural image datasets of 32×32 and 64×64 resolutions, and three medical datasets taken from MedMNIST (Yang et al., 2023) of 64×64 resolution (see Figure 8 for examples of MedMNIST images).

Table 1 presents the classification accuracy using both a linear probe and a MLP classifier. Across datasets, we observe substantial improvements with PMAE_{ocl} in linear probing compared to the standard MAE_{std}, with an average increase of 38% (+14 percentage points). Additionally, PMAE_{ocl} outperforms MAE_{ocl} by 20.3% (+9.5 percentage points). We observe similar trends with the randomized hyperparameter strategy: PMAE consistently outperforms MAE across all datasets, yielding an average performance increase of 29.9% (+5.4 percentage points), even when sub-optimal hyperparameters are used. These findings also extend to the non-linear evaluation setting, (see the middle part of Table 1). The bottom part of Table 1 presents the downstream performance in the fine-tuning setting. We continue to observe a performance gain brought by PMAE over MAE of 2.3% (+8.8 percentage points) on average. For two of the MedMNIST datasets we observe an equal and high ($> 98\%$) performance for both methods, showcasing that both approaches easily complete these tasks.

These empirical findings lead to several conclusions. We observe that the recommended masking of 75% of image patches is largely sub-optimal *across* datasets. Figure 5 reports an ablation study of the masking ratios for MAE and PMAE. Figure 5 (top) shows that, across all five datasets, a 75% masking ratio is sub-optimal. For PMAE, the masking ratio is a more stable hyperparameter. Figure 5 (bottom) shows that across all evaluated datasets we observe the best or near-optimal performance for PMAE at 20% of the variance masked. We also validate the empirical benefits brought by PMAE. Interestingly, we notice that PMAE without any hyperparameter tuning (PMAE_{rd}) outperforms MAE with optimum masking ratio in all but one case (i.e., CIFAR10). Finally, masking 20% of the variance in PMAE leads to substantial performance gains over the standard approach of using MAEs with a 75% masking ratio.

Figure 5 (right) shows the downstream performance over training epochs, where PMAE surpasses MAE as early as 200 epochs. Figures for other datasets are provided in Appendix A.2. Moreover, Figure 6 examines how downstream accuracy varies with different masking ratios. Our findings indicate that PMAE demonstrates similar or lower standard errors across masking ratios compared to MAE. Overall, these results highlight that the masking strategy of PMAE better aligns image reconstruction with image classification tasks compared to the MAE objective.

6. Understanding PMAE

In this section, we aim to provide more intuition as to why masking *components* rather than *image patches* leads to a more robust objective. In Section 2, we discuss the hypothesis under which MIM operates (Kong et al., 2023) and present an example failure case of spatial masking in Figure 3. We highlight how masking pixels can lead to a misalignment between the MIM objective and the learning of meaningful representations. If all patches covering an object are masked out, it is uncertain whether the remaining

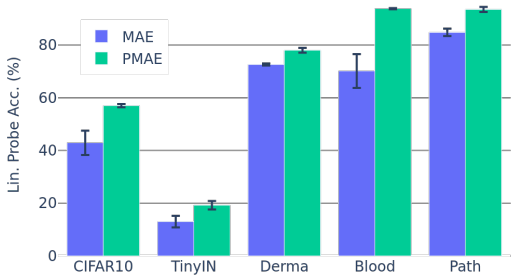

Figure 6. **Performance across masking ratios.** Average and standard error of the linear probe accuracy across masking ratios.

patches share any information with the object. On the contrary, if the masked-out information is redundant with the information carried by visible patches, it is likely that the information shared does not contain the object class but rather perceptual features (e.g., colors or textures).

Different from spatial masking, masking principal *components* leads to the removal of global image features, instead of only acting locally as in spatial masking. Figure 7 serves as an example highlighting the correspondence between principal component and perceptual features. In this example, the principal components with the highest eigenvalues capture the colors within the image while the bottom PCs highlight the edges. Early work in image processing (Turk & Pentland, 1991) has demonstrated this connection between an image's dominant modes of variation and its low spatial frequency components, providing further intuition for how information is partitioned in the space of PCs of natural images.

By removing a subset of principal components, PMAE prevents the removal of all the information characterizing an object and prevents redundant information to remain after masking. Instead, PMAE drops a set of unique image components. By taking advantage of the information partitioning in PCA, PMAE thereby mitigates MAE's failure cases, ultimately leading to increased accuracy. Although the potential of the principal component space for MIM (Balestriero & LeCun, 2024) or Image Denoising (Chen et al., 2024b) has been recently explored, our work is the first to propose an effective masking strategy that directly leverages PCA.

7. Related Work

Self-supervised learning. Self-supervised learning (SSL) leverages auxiliary tasks to learn from unlabeled data, often outperforming supervised methods on downstream tasks. SSL can be divided into two categories: discriminative and generative (Liu et al., 2021). Discriminative methods (Chen et al., 2020a; Caron et al., 2021) focus on enforcing invariance or equivariance between data views in the representation space, while generative methods (He et al., 2021; Bizeul et al., 2024) rely on data reconstruction from, often, corrupted observations. Though generative SSL historically lagged in performance, recent work has bridged the gap by integrating strengths from both paradigms (Assran et al., 2022; Dong et al., 2023; Oquab et al., 2023; Chen et al., 2024a; Lehner et al., 2023). Interestingly, recent discriminative methods employ cropping strategies to create distinct data views (Oquab et al., 2023; Assran et al., 2023), which is reminiscent of image masking. Balestriero & LeCun (2024) point out the misalignment between auxiliary and downstream tasks in reconstruction-based SSL and suggest novel masking strategies to help realign these objectives.

Masked Image Modelling. MIM extends the successful masked language modeling paradigm to vision tasks. Early methods, such as Context Encoder (Pathak et al., 2016), used a convolutional autoencoder to inpaint a central region of the image. The rise of Vision Transformers (ViTs) (Dosovitskiy et al., 2021) has driven significant advancements in MIM. BEiT (Zhou et al., 2021; Bao et al., 2022) combines a ViT encoder with image tokenizers (Ramesh et al., 2021) to predict discrete tokens for masked patches. SimMIM (Xie et al., 2022) simplifies the task by pairing a ViT encoder with a regression head to directly predict raw pixel values for the masked regions. MAE (He et al., 2021) introduces a more efficient encoder-decoder architecture, with a shallow decoder. MIM's domain-agnostic masking strategies have also proven effective in multi-modal tasks (Baevski et al., 2022; Bachmann et al., 2022).

Mask Design Strategies. A critical component of the masked image modeling paradigm is the design of effective masking strategies. Early MIM approaches have relied on random spatial masking techniques, such as masking out the central region of an image (Pathak et al., 2016), image patches (He et al., 2021; Xie et al., 2022), and blocks of patches (Bao et al., 2022). Inspired by advances in language modeling, recent efforts have explored semantically guided mask design. Li et al. (2021) use self-attention maps to mask irrelevant regions, while Kakogeorgiou et al. (2022) focus on masking semantically rich areas. Shi et al. (2022) design masks through adversarial learning, where the resulting masks resemble semantic maps, a concept extended by Li et al. (2022a) through progressive semantic region masking. Further advancing this direction, Wang et al. (2023) and Madan et al. (2024) introduce curriculum learning-inspired mask design methods. These methods often require additional training steps, components, or more complex objectives. More closely related to our work, Chang et al. (2022); Chen et al. (2024b) explore the use of pre-existing image representations for asked Image Modeling and image denoising. Chen et al. (2024b) introduce additive Gaussian noise to principal components as an alternative to the traditional Denoising Autoencoders. Chang et al. (2022) utilize masked token modeling by leveraging the discrete latent space of a pre-trained VQVAE to develop an image generation model.

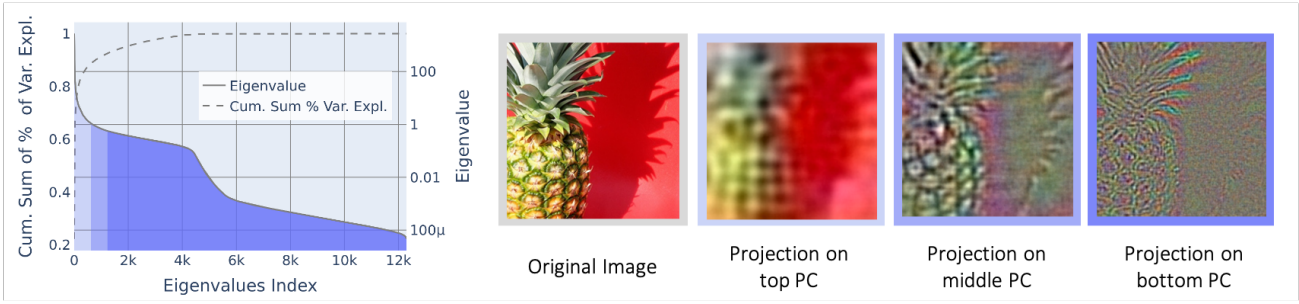

Figure 7. **From Principal Components to Spatial Features.** Overview of the spatial features associated with distinct regions of the principal component spectrum; Images depict the features captured by the top (light blue), middle (mild blue) and bottom (dark blue) PCs.

8. Discussion

In this work, we have investigated different masking strategies for Masked Image Modelling (MIM). To this end, we have introduced the Principal Masked Autoencoder (PMAE) as an alternative to spatial masking. PMAE is rooted in principal component analysis (PCA; Pearson, 1901; Hotelling, 1933), which is a widely used data-driven, *linear* and deterministic transformation. Unlike recent alternatives that require additional supervision, learnable components, or complex training pipelines (Li et al., 2021; 2022a; Kakogeorgiou et al., 2022; Li et al., 2022b), PMAE stays close to the core principles of MIM: the combination of a randomized masking strategy and an encoder-decoder architecture. Despite its simplicity, we demonstrate that PMAE yields substantial performance improvements over spatial masking on image classification tasks for CIFAR10, TinyImageNet and MedMNIST datasets. Further, in a PMAE, the masking ratio—typically a sensitive and difficult-to-tune hyperparameter in MIM—appears more robust and has a natural interpretation as the ratio of variance explained by the masked input.

Since PCA, or its generalisation kernel PCA (Schölkopf et al., 1997), is easily applicable to any data modality, our proposal of masking principal components is not specific to MIM. Instead, it can be viewed as a *general strategy* that should also be applicable to other types of modalities beyond images, as well as to other self-supervised learning (SSL) approaches. Indeed, data masking is commonly adopted in discriminative SSL methods. Whereas early approaches, such as SimCLR (Chen et al., 2020a) or MoCo (Chen et al., 2020b), relied on combinations of image transformations (e.g., color jitter, flips, crops, etc.) as data augmentation strategies, more recent state-of-the-art methods like DINO (Caron et al., 2021; Oquab et al., 2023) and I-JEPA (Assran et al., 2023) have shifted to relying solely on image cropping, which can be considered a type of masking. The integration of principal component masking instead of image cropping into such SSL pipelines constitutes a promising future direction of research.

In the present work, we have focused on PCA subspaces as meaningful masking spaces. However, our core idea of masking a *transformed* version of the data (rather than the *raw* data) can be viewed as laying the groundwork for other, more generic approaches to information masking for self-supervised representation learning. Moving beyond PCA, a natural extension would be to *learn* a suitable latent space in which the masking is performed. This route can potentially leverage recent theoretical insights (Kong et al., 2023) by more explicitly enforcing that the shared information between visible and masked-out latent components contains high-level latent variables that are most useful for the downstream tasks of interest. Other off-the-shelf transformations, such as the Fourier transform (Bracewell & Kahn, 1966), Wavelet transform (Daubechies, 1992), Kernel principal component analysis (Schölkopf et al., 1997), or Diffusion Maps (Coifman & Lafon, 2006), represent alternative candidate transformations. Future research should explore whether the properties of these spaces provide comparable or additional advantages over PCA. Preliminary results with non-linear transformations, namely Kernel PCA, presented in Appendix A.2.6, demonstrate performance gains over PMAE, motivating further exploration. A particularly appealing aspect of some of these methods (e.g., Fourier & Wavelet transforms and Diffusion Maps) is the use of fixed bases, which could eliminate the computational overhead of PCA—whose cost scales cubically with the data dimensionality—and improve scalability to larger datasets.

Our work demonstrates the potential of using PCA subspaces to guide image representation learning. Exploring alternative masking spaces with similar properties presents an exciting avenue for future research, offering the possibility of developing unsupervised tasks that encourage the learning of features more closely aligned with human perception.

Acknowledgments

We thank Randall Balestreiro, Carl Allen, Maximilian Dax, Georgios Kissas, and David Klindt for useful discussions and comments. Alice Bizeul is supported by an ETH AI Center Doctoral Fellowship. Alain Ryser is supported by the StimuLoop grant #1-007811-002 and the Vontobel Foundation.

References

Assran, M., Caron, M., Misra, I., Bojanowski, P., Bordes, F., Vincent, P., Joulin, A., Rabbat, M., and Ballas, N. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, 2022. URL http://arxiv.org/abs/2204.07141. [Cited on p. 7.]

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2301.08243. [Cited on p. 7 and 8.]

Bachmann, R., Mizrahi, D., Atanov, A., and Zamir, A. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/pdf/2204.01678. [Cited on p. 7.]

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2202.03555. [Cited on p. 1 and 7.]

Balestriero, R. and LeCun, Y. How learning by reconstruction produces uninformative features for perception. In *Forty-first International Conference on Machine Learning*, 2024. URL http://arxiv.org/abs/2402.11337. [Cited on p. 2, 4, and 7.]

Bao, H., Dong, L., Piao, S., and Wei, F. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. URL http://arxiv.org/abs/2106.08254. [Cited on p. 1, 3, and 7.]

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013. URL https://arxiv.org/pdf/1206.5538. [Cited on p. 3.]

Bizeul, A., Schölkopf, B., and Allen, C. A probabilistic model to explain self-supervised representation learning. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/pdf/2402.01399. [Cited on p. 7.]

Bracewell, R. and Kahn, P. B. The fourier transform and its applications. *American Journal of Physics*, 1966. [Cited on p. 8.]

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. URL http://arxiv.org/abs/2104.14294. [Cited on p. 7 and 8.]

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2202.04200. [Cited on p. 7.]

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020a. URL http://proceedings.mlr.press/v119/chen20j/chen20j.pdf. [Cited on p. 7 and 8.]

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b. URL https://arxiv.org/abs/2003.04297. [Cited on p. 8.]

Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., Han, S., Luo, P., Zeng, G., and Wang, J. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024a. URL https://link.springer.com/content/pdf/10.1007/s11263-023-01852-4.pdf. [Cited on p. 7.]

Chen, X., Liu, Z., Xie, S., and He, K. Deconstructing denoising diffusion models for self-supervised learning. *arXiv preprint arXiv:2401.14404*, 2024b. URL https://arxiv.org/pdf/2401.14404. [Cited on p. 2 and 7.]

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and computational harmonic analysis*, 2006. URL https://www.sciencedirect.com/science/article/pii/S1063520306000546. [Cited on p. 8.]

Daubechies, I. Ten lectures on wavelets. *Society for industrial and applied mathematics*, 1992. URL https://epubs.siam.org/doi/pdf/10.1137/1.9781611970104.fm. [Cited on p. 8.]

Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech- nologies*, 2019. URL https://arxiv.org/abs/1810.04805. [Cited on p. 1.]

Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N., and Guo, B. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL https://arxiv.org/pdf/2111.12710. [Cited on p. 1, 3, and 7.]

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the Tenth International Conference on Learning Representations.*, 2021. URL http://arxiv.org/abs/2010.11929. [Cited on p. 1, 4, 5, and 7.]

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sg d: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. URL https://arxiv.org/abs/1706.02677. [Cited on p. 13.]

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*, 2021. URL http://arxiv.org/abs/2111.06377. [Cited on p. 1, 2, 3, 4, 5, 6, 7, 12, 13, and 16.]

Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 1933. URL https://www.cis.rit.edu/~rlepci/Erho/Derek/Useful_References/Principal%20Components%20Analysis/Hotelling_PCA_part1.pdf. [Cited on p. 3 and 8.]

Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., and Komodakis, N. What to hide from your students: Attention-guided masked image modeling. In *ECCV*, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-20056-4_18. [Cited on p. 2, 7, and 8.]

Kong, L., Ma, M. Q., Chen, G., Xing, E. P., Chi, Y., Morency, L.-P., and Zhang, K. Understanding masked autoencoders via hierarchical latent variable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2306.04898. [Cited on p. 1, 2, 3, 6, and 8.]

Lehner, J., Alkin, B., Fürst, A., Rumetshofer, E., Miklautz, L., and Hochreiter, S. Contrastive tuning: A little help to make masked autoencoders forget. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. URL http://arxiv.org/abs/2304.10520. [Cited on p. 7.]

Li, G., Zheng, H., Liu, D., Wang, C., Su, B., and Zheng, C. SemMAE: Semantic-guided masking for learning masked autoencoders. *Advances in Neural Information Processing Systems*, 2022a. URL https://arxiv.org/abs/2206.10207. [Cited on p. 7 and 8.]

Li, J., Wang, Y., Zhang, X., Chen, Y., Jiang, D., Dai, W., Li, C., Xiong, H., and Tian, Q. Progressively compressed auto-encoder for self-supervised representation learning. In *The Eleventh International Conference on Learning Representations*, 2022b. URL https://openreview.net/pdf?id=8T4qmZbTkW7. [Cited on p. 8.]

Li, Z., Chen, Z., Yang, F., Li, W., Zhu, Y., Zhao, C., Deng, R., Wu, L., Zhao, R., Tang, M., and Wang, J. MST: Masked self-supervised transformer for visual representation. *Advances in Neural Information Processing Systems*, 2021. URL http://arxiv.org/abs/2106.05656. [Cited on p. 2, 7, and 8.]

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 2021. URL https://arxiv.org/pdf/2006.08218. [Cited on p. 7.]

Madan, N., Ristea, N.-C., Nasrollahi, K., Moeslund, T. B., and Ionescu, R. T. CL-MAE: Curriculum-learned masked autoencoders. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. URL https://arxiv.org/abs/2308.16572. [Cited on p. 7.]

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. URL https://arxiv.org/pdf/2304.07193. [Cited on p. 7 and 8.]

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. URL https://arxiv.org/pdf/1604.07379. [Cited on p. 1, 3, and 7.]

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901. URL https://doi.org/10.1080/14786440109462720. [Cited on p. 3 and 8.]

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International conference on machine learning*, 2021. URL http://proceedings.mlr.press/v139/ramesh21a/ramesh21a.pdf. [Cited on p. 7.]

Schölkopf, B., Smola, A., and Müller, K.-R. Kernel principal component analysis. In *International conference on artificial neural networks*, 1997. URL https://link.springer.com/chapter/10.1007/BFb0020217. [Cited on p. 8 and 16.]

Shi, Y., Siddharth, N., Torr, P., and Kosiorek, A. R. Adversarial masking for self-supervised learning. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. URL https://arxiv.org/abs/2201.13100. [Cited on p. 2 and 7.]

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 2021. URL https://arxiv.org/abs/2012.12877. [Cited on p. 5.]

Turk, M. and Pentland, A. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 1991. URL https://doi.org/10.1162/jocn.1991.3.1.71. [Cited on p. 7.]

Wang, H., Song, K., Fan, J., Wang, Y., Xie, J., and Zhang, Z. Hard patches mining for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2304.05919. [Cited on p. 7.]

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*, 2022. URL https://arxiv.org/abs/2111.09886. [Cited on p. 1 and 7.]

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 2023. URL https://www.nature.com/articles/s41597-022-01721-8. [Cited on p. 6 and 12.]

Yue, X., Bai, L., Wei, M., Pang, J., Liu, X., Zhou, L., and Ouyang, W. Understanding masked autoencoders from a local contrastive perspective. *arXiv preprint arXiv:2310.01994*, 2023. URL http://arxiv.org/abs/2310.01994. [Cited on p. 1.]

Zhang, Q., Wang, Y., and Wang, Y. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2210.08344. [Cited on p. 1 and 3.]

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/pdf/2111.07832. [Cited on p. 1, 3, and 7.]

A. Appendix

A.1. Experimental Setup

A.1.1. Datasets

CIFAR-10 is a widely used benchmark dataset containing 50,000 training and 10,000 validation 32×32 RGB images depicting 10 object classes, such as airplanes, cars, and animals.

TinyImageNet is a subset of the ImageNet dataset, containing 200 classes of 64×64 RGB images. It consists of 100,000 training images and 10,000 validation images, making it a challenging benchmark for classification tasks with more fine-grained object categories compared to CIFAR-10. Examples images are depicted in Figure 8

The MedMNIST (Yang et al., 2023) dataset is a collection of medical imaging datasets, each focusing on different types of biomedical data. In this work, three MedMNIST datasets are used:

BloodMNIST consists of 12,000 training and 1,700 validation 64×64 RGB images across 8 classes and represents microscopic images of blood cells, used for hematology classification tasks.

DermaMNIST contains 7,000 training and 1,000 validation 64×64 RGB images of skin samples, each part of one out of 7 types of skin diseases.

PathMNIST comprises 90,000 training and 10,000 validation 64×64 RGB images across 9 classes and depicts histopathological images of colorectal cancer tissue, aiding in classification tasks relevant to pathology.

We apply an equivalent data augmentation strategy to all datasets and for all learning objectives during training; Following He et al. (2021), our augmentation strategy consists of a random cropping followed by image resizing using bicubic interpolation. The scale of the random cropping is fixed to $[0.2, 1.0]$. We add horizontal flipping and we normalize images using each dataset's training mean and standard deviation; We keep these data augmentations during training and evaluation. We find that the use of data augmentations during the evaluation leads to a substantial performance drop for PMAE but we do keep these augmentations for fair comparisons. For all datasets and methods, we define image patches as patches of 8×8 pixels.

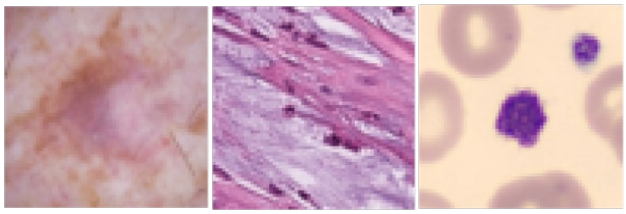

Figure 8. **MedMNIST datasets.** Example images from the (from left to right) DermaMNIST, PathMNIST, and BloodMNIST datasets used for image classification.

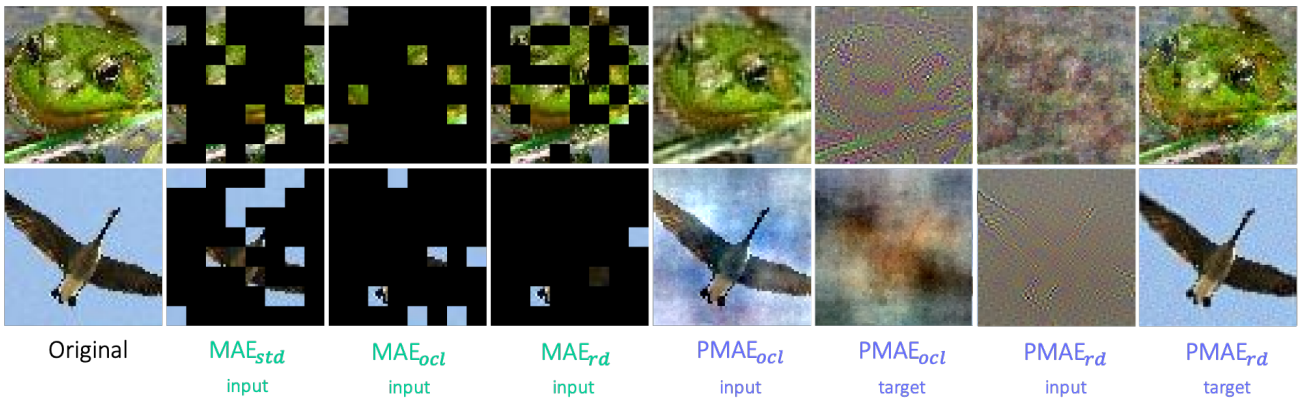

Figure 9. **Mask Design Strategies.** An overview of the different mask design strategies used in our experimental setup: spatial masking (*green*) and principal component masking (*blue*). *std* refers to the standard approach of masking out 75% of image patches, *ocl* denotes masking with the optimal masking ratio, *rd* represents a randomized strategy where the masking ratio is randomly sampled for each batch, and *target* refers to the reconstruction target.

A.1.2. Mask Design

Figure 9 depicts examples of the different masking strategies explored in Section 5. For MAE with *std* masking, the masking ratio is fixed to 75— 75% of patches are masked out. For MAE and PMAE with *ocl* masking, we tune the masking ratio to find the optimal ratio for each dataset and method. For *rd* masking, the masking ratio is sampled uniformly from the range [10,90] for each new batch of data during training.

A.1.3. Model Architecture

We train a tiny Vision Transformer encoder architecture (ViT-T) with image patch size 8×8 for all datasets (ViT-T/8). The specifics of this architecture can be found in Figure 10. Note that while for MAE, we keep the normalized pixel values of each masked patch as reconstruction targets (He et al., 2021), we find it to not have a clear positive impact for PMAE and hence enforce the reconstruction of the raw target values for PMAE.

config	**value**
hidden size	192
number of attention heads	12
intermediate size	768
norm pixel loss	Y/N
patch size	8×8

Figure 10. ViT-T/8 hyperparameters.

A.1.4. Training Hyperparameters

We train the ViT-T encoder-decoder architecture for 800 epochs with the hyperparameters found in Table 2. These hyperparameters are taken from (He et al., 2021). We use the linear lr scaling rule: lr = base lr×batchsize / 256 (Goyal et al., 2017). Note that for our oracle masking settings, we conduct ablation studies across a masking ratio range of $[10, 90]$.

A.1.5. Evaluation Hyperparameters

We evaluate the learned representation (i.e., `[CLS]` token) using a linear probe, multi-layer perceptron (MLP) classifier, and k-Nearest Neighbors algorithm on top of the frozen representation. The training samples of each dataset are used for training and the validation samples for testing. For linear and MLP probing experiments, we train the probes for 100 epochs following common practices (He et al., 2021). For the k-NN algorithm, we tune the number of neighbors in the range $[2, 20]$. More details regarding the linear probing evaluation hyperparameters can be found in Table 3. The same hyperparameters were used for the MLP probing. We also use the linear lr scaling rule: lr = base lr×batchsize / 256 (Goyal et al., 2017). For the fine-tuning setup, we also follow (He et al., 2021), we fine-tune the encoder and a linear probe for 100 epochs and resort to using the hyperparameters presented in Table 4. Note that for PMAE, we evaluate the approach on raw images and do not perform any filtering of principal components prior to evaluation.

A.1.6. Computational Resources

All training runs were conducted on a single NVIDIA GeForce RTX 3090/NVIDIA GeForce RTX 4090/Quadro RTX 6000 GPUs or NVIDIA TITAN RTX each of which has 24GB of VRAM. Figure 11 reports the time taken for 800 training epochs using a ViT-T/8 architecture on a Quadro RTX 6000 GPU for CIFAR10 and TinyImageNet with both MAE and PMAE.

config	**value**
batch size	512
base learning rate	0.00015
optimizer	AdamW [39]
betas (AdamW)	$\beta_1, \beta_2 = 0.9, 0.95$
learning rate	0.0003
warmup steps	40
weight decay	0.05

Table 2. Training hyperparameters.

config	**value**
batch size	512
base learning rate	0.1
optimizer	SGD [6]
betas (SGD)	0.9
learning rate	0.2
warmup steps	10
weight decay	0

Table 3. Linear probing hyperparameters.

config	**value**
batch size	512
base learning rate	0.001
optimizer	AdamW [39]
betas (AdamW)	$\beta_1, \beta_2 = 0.9, 0.99$
learning rate	0.002
warmup steps	5
weight decay	0.5

Table 4. Fine-tuning hyperparameters.

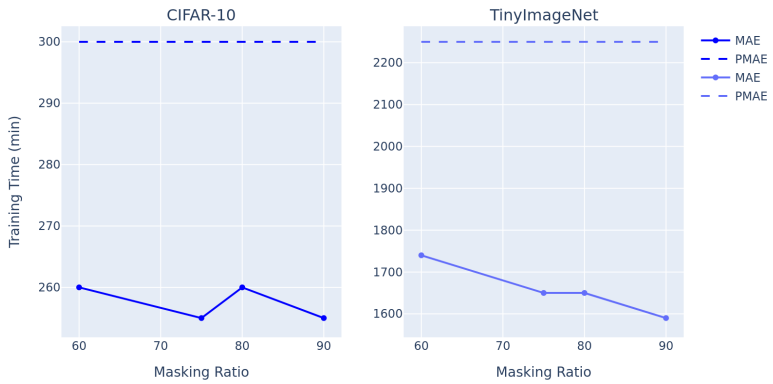

Figure 11. **Training time.** We report the training time in minutes for 800 training epochs using a ViT-T/8 architecture. For standard MAE we report numbers for various masking ratios.

A.2. Additional Results

A.2.1. Reconstruction of Masked Information

Figure 12 depicts the output of the decoder networks after 1000 training epochs for both MAE and PMAE. Interestingly, we observe that for PMAE, the model is able to well estimate the masked out principal components.

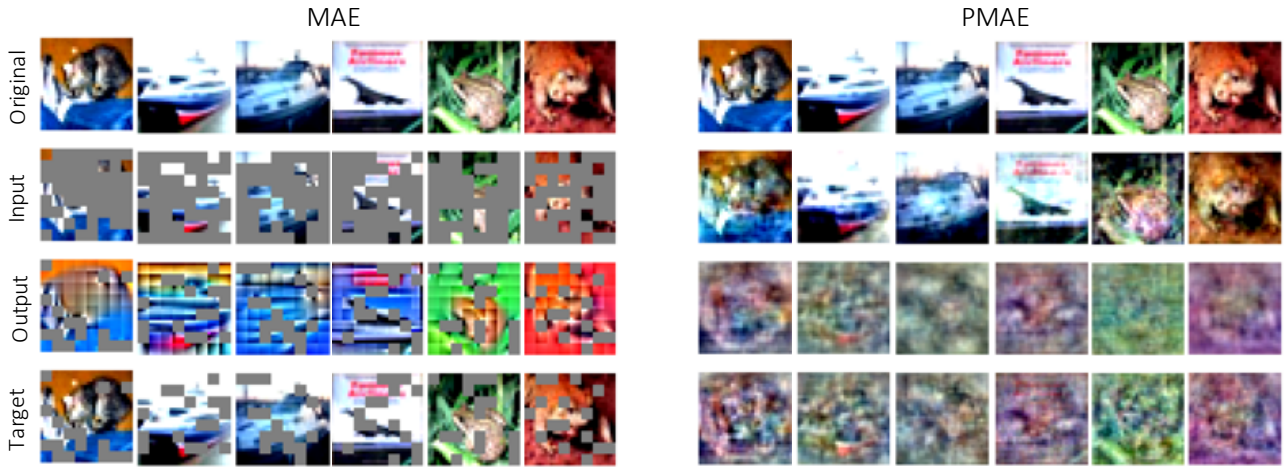

Figure 12. **Reconstruction of masked information.** Example of original images, model input and output as well as the target for MAE (left) and PMAE (right) for CIFAR10. Training was performed for 1000 epochs with masking ratio of 75% and 10% for MAE and PMAE respectively. Reconstructions were obtained using the validation dataset.

A.2.2. Training Dynamics

Figure 17 displays the linear probe accuracy for varying training epoch checkpoints. Similar to Figure 5 (right) we observe that PMAE after 200 epochs outperforms MAE after 800 epochs.

A.2.3. k-Nearest Neighbor Evaluation

Table 5 shows the classification accuracy for CIFAR10, TinyImageNet, and MedMNIST datasets using a k-NN classifier in place of a linear or MLP probe as presented in Section 5 and Appendix A.2.4.

A.2.4. Reconstructing in pixel vs. principal component space

We further investigate the impact of the domain (i.e., pixel vs. pc space) in which the reconstruction error is minimized on downstream performance. In Figure 18, we present an alternative to Figure 2 in which the training objective receives a set of pixels in place of principal components. In Figure 18, the masked out principal components are projected back into pixel space and the decoder's output is kept as is. The training objective (Equation (A.1)) then minimizes the Euclidean distance

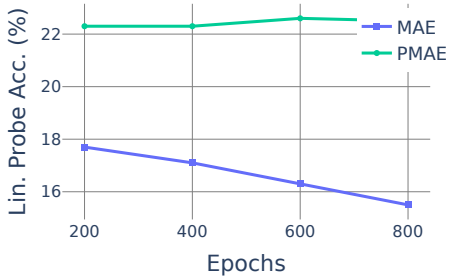

Figure 13. TinyImageNet

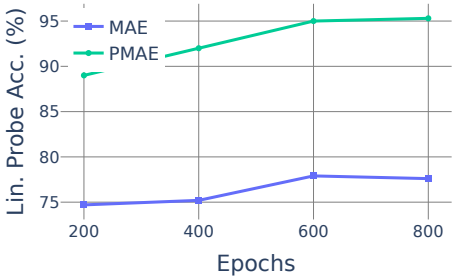

Figure 14. BloodMNIST

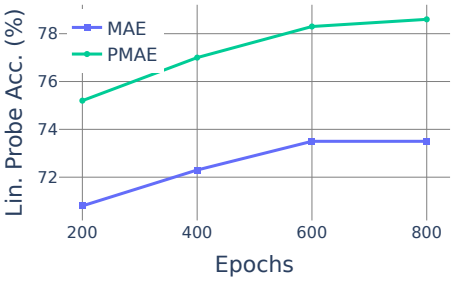

Figure 15. DermaMNIST

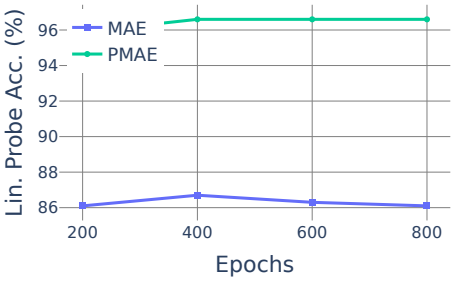

Figure 16. PathMNIST

Figure 17. **Performance Curves.** Linear probe accuracy (%) for TinyImageNet and MedMNIST datasets across training epochs. PMAE, after 200 epochs, outperforms MAE after 800 epochs.

between the ground truth masked principal components projected back to pixel space and the decoder's output. Instead in Figure 2 and Equation (3.1), the training objective minimizes the Euclidean distance between the ground truth and the predicted masked principal components. Equation (A.1) becomes a modified version of Equation (3.1):

$$\mathcal{L}_{\text{PMAE}}(\mathbf{x}, \mathbf{m}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \left\| \left(g_{\boldsymbol{\theta}} \circ f_{\boldsymbol{\phi}} \circ t^{-1} \left(\mathbf{m} \odot t\left(\mathbf{x} \right) \right) \right) - t^{-1} \left((\mathbf{1} - \mathbf{m}) \odot t(\mathbf{x}) \right) \right\|_2^2, \tag{A.1}$$

Table 7 presents the downstream image classification performance achieved when training representations with Equation (A.1). In particular, we report results obtained using a linear and MLP probe in the standard, oracle and randomized masking settings. Despite lower performance gains as the ones presented with Equation (3.1) in Table 1, PMAE consistently outperforms MAE across all five datasets, demonstrating substantial improvements. In Table 1 we report an average performance gain of 9.6 percentage points across datasets over the MAE baseline, while Table 7 reports an average performance gain of 6.6 percentage points. These findings also support our claims that the space of principal components constitutes a meaningful masking space for Masking Image Modelling learning paradigms.

A.2.5. Masking Ratio Ablation

In Appendix A.2.6, we perform an ablation across the masking ratio for representations learned with Equation (A.1). We draw similar conclusions to the ones drawn from Appendix A.2.6: the optimal masking ratio across datasets lies between 10 and 20% of the variance masked. Above these ratios, we observe a performance drop across datasets. Appendix A.2.6 also shows the impact of the use of data augmentations during training of the linear probe on the downstream accuracy. We see a significant increase in performance for PMAE when dropping data augmentations for the training of the linear probe but still resort to keeping these augmentations for the main results presented in Table 1 to ensure a fair comparison with the MAE.

A.2.6. Beyond PCA

Our work shows evidence that PCA offers a meaningful masking space. In Section 6, we motivate our choice by observing that principal components capture global rather than local features of an image. In this section, we go beyond PCA and explore non-linear matrix factorization methods as a proof of concept for future research. In particular, we explore Kernel

Table 5. **Evaluation with k-NN.** k-Nearest Neighbors top-1% accuracy for CIFAR10, TinyImageNet and MedMNIST for masking in pixel (MAE) and principal component (PMAE) space with the standard 75% masking ratio (std), oracles (ocl) and randomized (rd) masking ratios. $\star$ and $\star\star$ refer to representations trained with Equation (3.1) and Equation (A.1) respectively.

		CIFAR10	TinyImageNet	DermaMNIST	BloodMNIST	PathMNIST
k-NN	MAE_{std}	38.3	10.0	71.1	65.7	92.1
	MAE_{ocl}	47.6	12.5	69.9	73.6	94.6
	$\text{PMAE}^{\star}_{\text{ocl}}$	**55.3**	**14.2**	**76.6**	**93.1**	**99.6**
	$\text{PMAE}^{\star\star}_{\text{ocl}}$	48.1	9.6	74.7	84.5	99.1
	MAE_{rd}	40.3	7.6	71.6	82.7	**96.0**
	$\text{PMAE}^{\star}_{\text{rd}}$	41.8	**11.2**	**72.1**	**83.6**	95.5
	$\text{PMAE}^{\star\star}_{\text{rd}}$	**49.6**	9.5	70.6	76.0	94.8

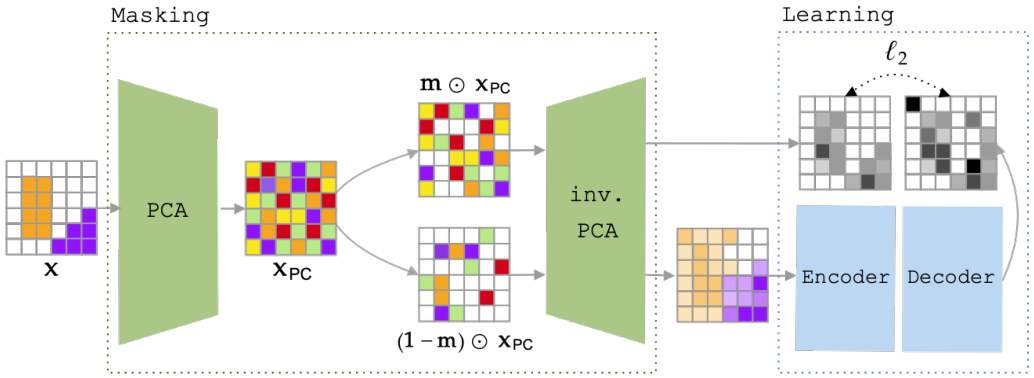

Figure 18. **Overview of the Principal Masked Autoencoder (PMAE) with reconstruction targets in the observation space.** A Principal Masked Autoencoder (PMAE) differs from a vanilla MAE by performing the masking in the space of principal components $\mathbf{x}_{\text{PC}} = \text{PCA}(\mathbf{x})$ rather than in the pixel space. The masked and visible principal component $\mathbf{m} \odot \mathbf{x}_{\text{PC}}$ and $(\mathbf{1}-\mathbf{m}) \odot \mathbf{x}_{\text{PC}}$ are then projected back into the space of pixels and serve as the reconstruction target and input for an encoder-decoder architecture, respectively.

PCA (Schölkopf et al., 1997) with a Radial Basis Function (RBF) kernel with the kernel coefficient set to 3.10^{-4}. In kernel PCA, the spectral decomposition is performed not on the data itself but rather on a modified version of it: the standardized data is mapped to a high-dimensional space via a non-linear kernel function.

In Table 6, we present results on the CIFAR10 dataset and show the image classification accuracy using a linear and MLP probe. We compare a vanilla MAE with our PMAE and KMAE which relies on Kernel PCA for optimal masking ratios (30% of the variance is masked out). For KMAE, we use the setting presented in Appendix A.2.4 and use Equation (3.1) as the training objective.

Table 6. Linear and MLP probe accuracy for CIFAR10 for random masking in pixel (MAE), in principal component space (PMAE), and in Kernel PCA space (KMAE) with the standard 75% masking ratio (std) and oracles (ocl). * refers to ours.

	MAE_{std}	MAE_{ocl}	$\text{PMAE}^{*}_{\text{ocl}}$	$\text{KMAE}^{*}_{\text{ocl}}$
Linear	41.7	50.7	59.0	**64.1**
MLP	34.0	55.2	64.1	**68.6**

The results reveal a significant performance improvement when employing a non-linear image transformation. KMAE achieves an average gain of 13.3 and 5 percentage points compared to the standard Masked Autoencoder (He et al., 2021) and PMAE, respectively. Although these findings are preliminary and based on a single mid-scale dataset, they highlight the potential of non-linear transformations and further emphasize the value of spectral decomposition as a meaningful for masked image modeling paradigms.

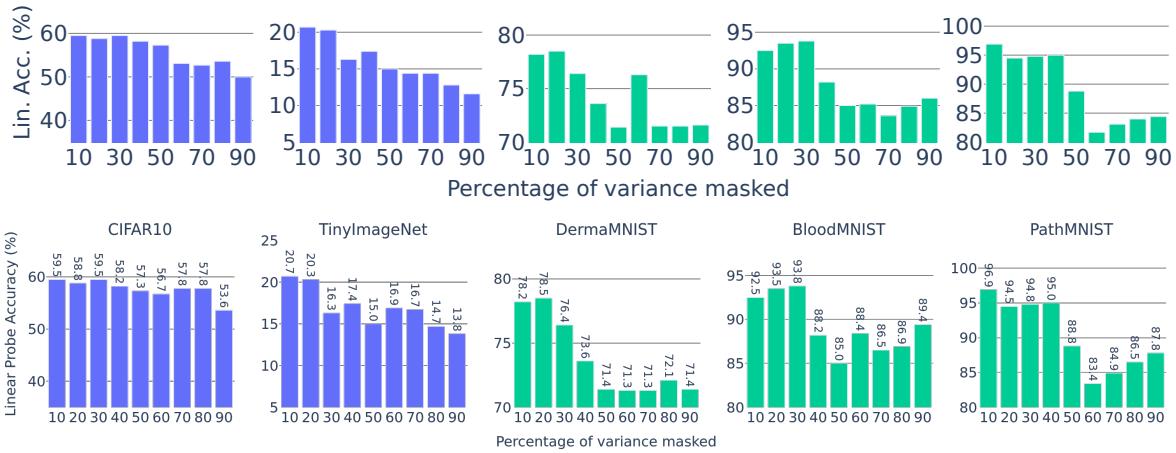

Figure 19. **Impact of the Masking Ratio.** (*top*) PMAE trained with Equation (A.1) linear probing accuracy for varying masking ratios. We observe a close-to-optimal performance across datasets for 10 to 20% of the variance masked. *(bottom)* PMAE trained with Equation (A.1) linear probing accuracy *without* data augmentations at evaluation. We observe a significant performance improvement without data augmentations.

Table 7. Linear & MLP probe and fine-tuning top-1% accuracy for CIFAR10, TinyImageNet and MedMNIST datasets for random masking in pixel (MAE) and principal component (PMAE) space with reconstruction target in pixel space with the standard 75% masking ratio (std), oracles (ocl) and randomized masking ratios (rd). Representations are learnt from Equation (A.1). * refers to ours.

		CIFAR10	TinyImageNet	DermaMNIST	BloodMNIST	PathMNIST
Linear	MAE_{std}	41.7	11.5	72.4	73.4	83.4
	MAE_{ocl}	50.7	15.5	73.7	78.6	86.4
	$\text{PMAE}^{*}_{\text{ocl}}$	**55.1**	**17.4**	**77.4**	**91.0**	**97.0**
	MAE_{rd}	41.9	7.5	72.4	83.2	85.6
	$\text{PMAE}^{*}_{\text{rd}}$	**56.0**	**15.1**	**74.5**	**85.9**	**87.5**
MLP	MAE_{std}	34.0	15.5	72.2	68.6	92.6
	MAE_{ocl}	55.2	**22.2**	74.4	75.8	95.1
	$\text{PMAE}^{*}_{\text{ocl}}$	**61.5**	22.1	**79.6**	**91.0**	**98.8**
	MAE_{rd}	38.5	11.6	66.9	70.6	95.7
	$\text{PMAE}^{*}_{\text{rd}}$	**62.2**	**19.5**	**75.3**	**84.4**	**97.0**