Reducing data in visual AI

Assessing the Data Efficiency of Masked Autoencoders in Resource-Constrained Environments

Dimo Terziev[1]

Supervisor(s): Jan van Gemert[1]**, Petter Reijalt**[1]**, Alex Manolache**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Dimo Terziev
Final project course: CSE3000 Research Project
Thesis committee: Jan van Gemert, Petter Reijalt, Alex Manolache, Mitchell Olsthoorn

Abstract

Visual foundation models based on Vision Transformers often depend on large datasets and substantial computational resources, limiting their accessibility for resource-constrained research settings. This paper investigates the data efficiency of Masked Autoencoders (MAE) by studying how pre-training dataset size and mask ratio affect downstream representation quality. A MAE model is pre-trained on nested subsets of the same dataset ranging from 1k to 100k images, using different mask ratios, and then evaluated on a different downstream task dataset. The results show that MAE learns transferable representations even from small unlabeled datasets, with downstream accuracy increasing steadily as more pre-training data is used. The experiments also show that the optimal masking difficulty depends on the data regime: lower masking improves validation accuracy for the smallest subsets, while the original 75% MAE mask ratio becomes stronger as the dataset size increases. These findings suggest that mask ratio should not be treated as a fixed default in MAE training. Instead, reducing the mask ratio can improve data efficiency when pre-training data is limited, while higher masking remains effective when more visual variation is available.

1 Introduction

Training reusable vision models often requires more data and compute than small research groups can realistically access. This problem is especially visible in visual foundation models, where a model is first trained on a broad image dataset and later reused for several downstream tasks [1]. In computer vision, this development is strongly connected to the Vision Transformer (ViT) architecture [2]. A Vision Transformer divides an image into small patches and processes these patches with a transformer, a neural network architecture originally developed for sequence data. This approach has shown strong performance when trained at sufficient scale, but it often depends on very large amounts of data and compute.

This dependency on large datasets creates an important problem for research. If competitive visual models can only be trained using massive datasets and large computational resources, then the development of state-of-the-art models becomes limited to highly resourced organizations. This makes it more difficult for smaller institutions, students, and independent researchers to study, reproduce, or adapt these models. It also creates ethical concerns, because large datasets are often difficult to inspect fully and may contain bias, copyright issues, or unfair demographic distributions [1]. Therefore, reducing the dependency on huge datasets is not only a technical goal, but also a way to make visual AI research more accessible and responsible.

Self-supervised learning offers one way to reduce the dependency on large amounts of data. Instead of requiring a human label for every training image, self-supervised methods create a training signal from the data itself. Masked Autoencoders (MAE) are a self-supervised method for Vision Transformers in which a large fraction of image patches is hidden and the model is trained to reconstruct the missing patches [3]. After this pre-training stage, the encoder can be reused as a representation model for downstream tasks. This makes MAE a useful method for studying data efficiency, because it allows visual representations to be learned from unlabeled images.

However, the data efficiency of MAE may depend on more than the number of available images. MAE has a method-specific hyperparameter - the mask ratio, which controls how many image patches are hidden during pre-training, and therefore directly controls the difficulty of the reconstruction task. This makes it different from general training choices such as batch size or learning rate, because it changes the learning signal produced by each image. A higher mask ratio gives the model less visible context and forces it to reconstruct more missing content. The original MAE setup uses a 75% mask ratio, but this choice was developed for large-scale pre-training [3]. When the amount of pre-training data is small, the same masking difficulty may not be optimal. A lower mask ratio may give the model more useful information from each image and therefore improve learning in small-data settings.

This paper studies how pre-training dataset size and mask ratio affect MAE representation quality. The encoder is pre-trained on nested subsets of the same dataset and evaluated with linear probing. In linear probing, the pre-trained backbone is frozen and only a linear classifier is trained on the downstream labels [4; 3], so the evaluation measures what was learned during pre-training rather than how well the full model can adapt during fine-tuning [5]. The experiments therefore focus on the representation learned by MAE. Because different dataset sizes may require different numbers of updates before the learned representation stops improving, checkpoint selection is based on a plateau criterion rather than on a fixed number of epochs.

The main research question is:

> How do pre-training dataset size and mask ratio affect the data efficiency of Masked Autoencoders?

This question is divided into two subquestions:

1. How does downstream linear-probe accuracy change as the amount of pre-training data increases?
2. How does the MAE mask ratio affect validation accuracy at different pre-training dataset sizes?

The experiments show that MAE learns transferable representations even with small pre-training subsets, and that downstream accuracy increases as more unlabeled pre-training data is used. They also show that the best mask ratio depends on the data regime. Lower masking works better for the smallest subsets, while the original 75% mask ratio becomes stronger when more pre-training data is available. The main contribution of this paper is an empirical analysis of MAE data efficiency that connects dataset size, masking difficulty, and downstream representation quality in a constrained training setting.

2 Related Work

Vision transformers can achieve high image-classification accuracy, but this often depends on large training datasets. This data requirement follows from the way vision transformers process images with fewer built-in image assumptions than convolutional models [2]. Training recipes for vision transformers have reduced part of this data requirement, but such recipes still depend on enough labeled or unlabeled data to learn useful visual structure [6]. Large foundation-model pipelines make the same problem more practical: when useful models require large datasets, smaller research groups have less ability to reproduce, inspect, and adapt them [1]. For this reason, the amount of data needed for useful ViT representations is the main concern behind this paper.

Self-supervised pre-training reduces the need for labels by learning from the images themselves. Masked image modeling is one form of self-supervised pre-training, where part of an image is hidden and the model learns to predict the missing content [7; 8]. MAE follows this masked-reconstruction approach with an encoder–decoder design: the encoder receives only the visible image patches, while the decoder reconstructs the missing patches [3]. This design makes MAE relevant for constrained training, because the encoder processes fewer tokens during pre-training while still learning from unlabeled images. The key question for this paper is therefore not whether MAE can learn visual representations in general, but how well MAE keeps learning when the pre-training dataset becomes small.

Prior MAE studies show that masked reconstruction can be effective, but the strongest evidence often comes from larger or fixed pre-training settings. The original MAE work reports strong results when the method is trained with a high mask ratio and a large pre-training setup [3]. Small-scale MAE studies show that the same idea can also be adapted to smaller models and smaller image datasets, but their results depend on choices such as patch size, reconstruction target, and masking strategy [9; 10]. These studies are useful for choosing MAE as the method to study, but they do not fully answer how MAE accuracy changes across a controlled range of pre-training dataset sizes. This paper addresses that gap by measuring the learned representation after pre-training on progressively smaller dataset fractions.

The mask ratio is closely tied to MAE data efficiency because the mask ratio changes the amount of information available during reconstruction. A high mask ratio leaves few visible patches, which can reduce encoder compute and force the model to learn broader image structure [3]. A lower mask ratio leaves more visible patches, which can make the reconstruction task easier and may give a more stable learning signal when few images are available. Prior masked-reconstruction work often treats the mask ratio as a training choice for a fixed data setting, rather than as a factor that may interact with dataset size [3; 9]. This paper therefore studies the mask ratio together with dataset size, while using a shared checkpoint-selection protocol to account for differences in training length across dataset sizes.

Linear probing is a suitable evaluation protocol when the goal is to compare learned representations under a limited compute budget. Since only a linear classifier is trained on the downstream labels, linear probing is much faster to compute than full fine-tuning, because only a small number of parameters are updated [4]. Full fine-tuning can improve downstream accuracy, but it also updates the backbone and makes the effect of pre-training harder to isolate [5]. Linear probing therefore matches the goal of this paper, because it provides a quick downstream evaluation while still showing how dataset size and mask ratio affected the representation learned during MAE pre-training.

3 Methodology

This section describes how MAE pre-training and linear probing are used to measure representation quality. The method has two stages. First, an encoder is pre-trained without labels by reconstructing missing image patches. Second, the pre-trained encoder is frozen and evaluated with a linear classifier. This separation is important because the paper studies the representation learned during pre-training, not the ability of the full model to adapt during fine-tuning.

3.1 MAE Pre-training

MAE pre-training learns visual representations by reconstructing hidden parts of an image. An input image is first divided into non-overlapping patches. A fixed fraction of these patches is then masked. The remaining visible patches are passed to the encoder, while the masked patches are removed from the encoder input. The mask ratio controls how much of the image is hidden, so the mask ratio also controls the difficulty of the reconstruction task. Figure 1 shows the MAE pre-training pipeline. The figure highlights the role of the mask ratio: increasing the mask ratio removes more patches from the encoder input, while decreasing the mask ratio leaves more visible context for reconstruction.

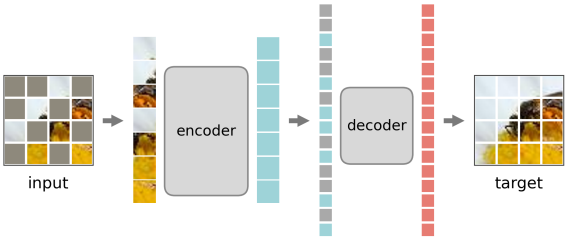

Figure 1: Masked Autoencoder pre-training pipeline, based on the method from He et al. [3]. An input image is split into patches, a subset of patches is masked, and only the visible patches are processed by the encoder. A lightweight decoder then reconstructs the missing patches from the encoded visible patches and learned mask tokens. In this paper, the mask ratio is varied because it changes how much context is available to the encoder during reconstruction.

The encoder produces a representation from only the visible patches. After the encoder, a lightweight decoder receives the encoded visible patches together with learned mask tokens. The mask tokens mark the locations where patches were removed. Positional embeddings are added so that the

decoder can use the original patch locations. The decoder then predicts the pixel values of the missing patches. After pre-training, the decoder is discarded and only the encoder is kept for downstream evaluation.

The reconstruction loss compares the predicted patches with the original image patches. Let x_i be the original pixel values of patch i, let $\hat{x}_i$ be the reconstructed pixel values of the same patch, and let $\mathcal{M}$ be the set of masked patches. The MAE reconstruction loss is

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|x_i - \hat{x}_i\|_2^2 . \tag{1}$$

Equation 1 trains the model to predict the missing image content from the visible context. This objective is useful for this paper because the learning signal comes from the image itself, which makes MAE suitable for studying unlabeled pre-training under different data budgets.

3.2 Representation Evaluation

The quality of the pre-trained encoder is evaluated with linear probing. Linear probing freezes the encoder and trains only a linear classifier on top of the encoder representation. Freezing the encoder prevents the downstream task from changing the learned representation. The resulting accuracy therefore gives a direct estimate of how useful the pre-trained representation already is. Figure 2 illustrates the linear-probing pipeline used to evaluate the frozen MAE encoder.

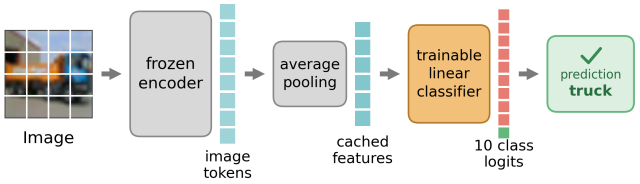

Figure 2: Linear-probing evaluation pipeline. A downstream image is passed through the frozen pre-trained encoder, producing patch-level image tokens. These tokens are aggregated with average pooling into a fixed feature vector, and only the linear classifier is trained to predict the downstream class. This setup evaluates the representation learned during MAE pre-training without updating the encoder.

Let $f_\theta(x)$ be the frozen encoder representation of an input image x. The encoder parameters θ are fixed during linear probing. Let W and b be the trainable weight matrix and bias vector of the linear classifier. The classifier produces logits

$$z = W f_\theta(x) + b. \tag{2}$$

In Equation 2, z contains one score for each downstream class. Using these scores, the linear classifier is trained by trying to minimize the cross-entropy loss. Let C be the number of classes, let y_c be the target label for class c, and let p_c be the predicted probability for class c after applying the softmax function to z. The loss is

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \log p_c. \tag{3}$$

Only W and b are updated when minimizing Equation 3. This evaluation protocol is therefore suitable for this paper, since many pre-training settings must be compared under a limited compute budget, which is more difficult when there are more parameters that need to be updated.

3.3 Studied Factors

The method is used to study two factors that affect MAE data efficiency. The first factor is the amount of pre-training data. Changing the amount of pre-training data changes how many images the MAE objective can learn from before downstream evaluation. The second factor is the mask ratio. Changing the mask ratio changes how much visual context is available during reconstruction.

These two factors are connected. With more pre-training data, the model can learn from more visual variation, so a harder reconstruction task may still provide a useful learning signal. With less pre-training data, a high mask ratio may remove too much information from each image, which can make the reconstruction task less helpful. For this reason, the experiments evaluate dataset size and mask ratio together rather than treating the mask ratio as a fixed setting.

4 Experimental Setup

This section describes the experimental choices needed to understand the results. Complete configuration tables for MAE pre-training and downstream linear probing are provided in Appendix A. Across the main experiments, only the pre-training dataset size and mask ratio were varied.

4.1 Datasets

The dataset used for MAE pre-training was Tiny ImageNet. It was chosen, because this dataset is small enough for repeated pre-training runs on limited hardware, but large enough to study how representation quality changes with dataset size. Its 64×64 resolution also keeps the patch-based reconstruction task meaningful while being cheaper than full-resolution ImageNet.

To study data efficiency, the Tiny ImageNet training set was split into nested subsets of 1k, 2k, 4k, 8k, 16k, 32k, 64k, and 100k images. Each smaller subset was contained in all larger subsets. This makes the comparison between dataset sizes more controlled, because increasing the dataset size only adds images instead of replacing earlier ones. The Tiny ImageNet validation set was used only for monitoring and checkpoint selection during pre-training, but was not used for the pre-training itself.

CIFAR-10 was the dataset chosen for the final downstream evaluation. This gives an out-of-domain test, because the encoder is pre-trained on Tiny ImageNet and evaluated on a different classification dataset. All CIFAR-10 images were resized to 64×64 and normalized with ImageNet statistics.

4.2 Model and Pre-training Configuration

All experiments used a ViT-Tiny/8 encoder with hidden size 192. The MAE decoder used hidden size 192, 4 transformer layers, 3 attention heads, and intermediate size 768. These values were chosen to keep the decoder lightweight while matching the encoder representation size. The hidden size and number of attention heads keep the decoder compatible with the ViT-Tiny/8 encoder, while the smaller depth limits the extra compute used only for reconstruction. This is important in this study because the decoder is discarded before evaluation, so a large decoder would increase pre-training cost without directly improving the downstream model.

The pre-training configuration mostly followed the original MAE recipe [3]. The only difference in hyperparameter values was the batch size, which was chosen to be 500, because it divides all Tiny ImageNet subset sizes exactly, while being the largest value that fit the GPU on the local PC used for testing.

The mask ratio was varied across three values: 62.5%, 75%, and 87.5%. The 75% setting matches the original MAE setup and leaves 16 visible patches per image in this configuration [3]. The 62.5% setting leaves 24 visible patches and tests whether smaller datasets benefit from an easier reconstruction task. The 87.5% setting leaves 8 visible patches and tests whether stronger masking remains useful when less visual context is available.

4.3 Plateau-based Training Protocol and Checkpoint Selection

Pre-training used a warmup–stable–decay schedule for plateau-based checkpoint selection [11]. This schedule was chosen for the experimental protocol, not as a factor in the research question. The goal of the schedule was to make runs extendable: after warmup, the stable phase keeps the learning rate constant, so training can continue while the monitored representation quality still improves. This is useful when different pre-training dataset sizes may require different numbers of updates before plateauing. In contrast, a cosine schedule requires the final training length to be fixed in advance, which would make it harder to distinguish a true representation-quality plateau from a plateau caused by a decayed learning rate [12].

The schedule starts with a 40-epoch warmup, following the original MAE setup [3]. After warmup, the learning rate stays constant during the stable phase. This constant phase is useful for this study because the run length is not fixed in advance: training can continue while the monitored representation still improves, which takes different time on the different dataset splits. The stable phase ends when validation accuracy plateaus or when the run reaches the 100k-step compute limit.

After the stable phase, the learning rate was decayed during a cooldown phase. Different cooldown lengths have been studied for constant learning-rate training, with useful values reported around 10–20% of total steps [12]. The cooldown in this paper was set to 10% of the total run length.

During the stable phase, a monitoring probe was trained every 40 epochs. The monitoring probe used the current Tiny ImageNet subset for training and the Tiny ImageNet validation set for evaluation. Training stopped when no new best validation accuracy was found for 10 consecutive monitoring probes, which corresponds to 400 epochs without improvement. For final CIFAR-10 evaluation, the selected checkpoint was the one with the highest Tiny ImageNet validation accuracy, not necessarily the last checkpoint of the run.

4.4 Downstream Linear-Probe Evaluation

The final representation was evaluated with a linear probe on CIFAR-10. The probe input was produced with global average pooling over the encoder patch tokens, following the MAE linear-probing setup [3]. The probe head used $\mathrm{BatchNorm1d}(192)$ with `affine=false`, followed by $\mathrm{Linear}(192, 10)$. Batch normalization was introduced to normalize activations during training [13], and the non-affine version is used in MAE linear probing to normalize feature magnitudes without adding trainable affine parameters [3].

4.5 Compute Environment

All final experiments were run in a RunPod container with one NVIDIA RTX A4500 GPU. The instance used 12 vCPUs from an AMD EPYC 7352 24-Core Processor, 62 GB of system memory, and 20 GB of container disk. The software environment was `runpod/pytorch:2.4.0-py3.11-cuda12.4.1-devel-ubuntu22.04`. Each MAE pre-training configuration was run with seeds 0 and 42. For each selected pre-trained checkpoint, the downstream CIFAR-10 linear probe was trained with seed 0. The reported pre-training results therefore average two MAE runs, while the downstream probe uses a fixed probe seed to isolate the effect of the pre-trained encoder.

5 Results

This section presents the experimental results in two steps. First, the mask-ratio experiment tests whether the best masking difficulty changes with pre-training dataset size. Second, the downstream experiment evaluates how CIFAR-10 accuracy changes when MAE is trained with the selected mask-ratio rule. The first experiment uses Tiny ImageNet validation accuracy for checkpoint monitoring, while the second experiment uses CIFAR-10 test accuracy after downstream linear probing.

5.1 Experiment 1: Which mask ratio works best at different dataset sizes?

Question. How does the MAE mask ratio affect validation accuracy when the amount of pre-training data changes?

Figure 3 shows the effect of mask ratio on Tiny ImageNet validation accuracy. The plot reports the absolute difference in percentage points relative to the original MAE mask ratio of 75%. This makes the comparison easier to read than raw validation accuracy, because the validation accuracy changes strongly with dataset size.

The lower mask ratio of 62.5% gives the best result on the smallest subsets. At 1k images, it is slightly above the 75% mask ratio, while the 87.5% mask ratio is clearly worse. The

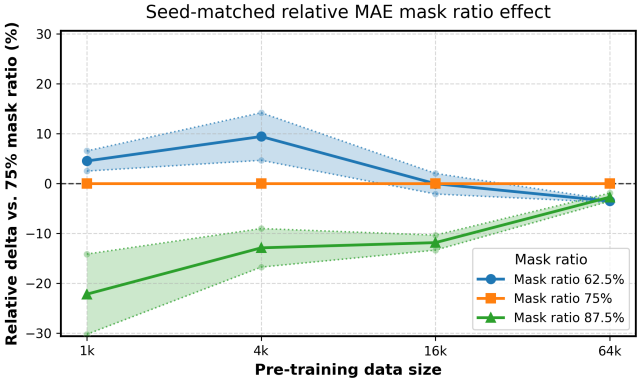

Figure 3: Effect of mask ratio on Tiny ImageNet validation accuracy. The y-axis shows the absolute difference in percentage points compared with the original MAE mask ratio of 75% at the same dataset size. Positive values mean that a mask ratio performs better than 75%, while negative values mean that it performs worse. Lower masking improves validation accuracy for the 1k and 4k subsets, the mask ratios are close at 16k, and the original 75% mask ratio is strongest at 64k.

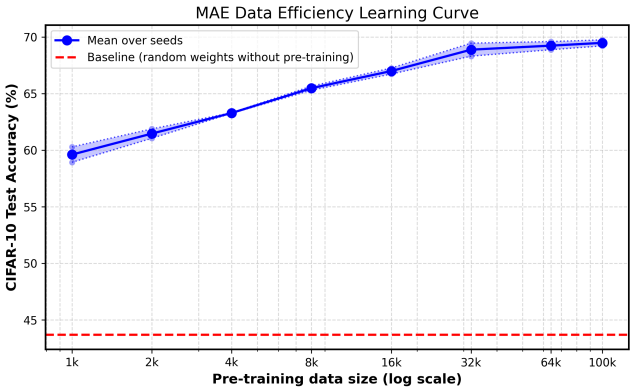

Figure 4: MAE data-efficiency learning curve. The figure shows CIFAR-10 top-1 test accuracy after MAE pre-training on different Tiny ImageNet subset sizes. The selected mask-ratio rule from Figure 3 is used: lower masking for smaller subsets and the original 75% mask ratio for larger subsets. Accuracy increases as more pre-training data is used, and all pre-trained models are well above the random-weight baseline. The smaller gain after 32k images should be interpreted together with the training limit, since the largest pre-training runs reached the maximum update budget before plateauing.

difference becomes larger at 4k images: 62.5% masking improves over 75% by about 1.3 percentage points, while 87.5% masking is about 2.0 percentage points lower. At 16k images, 62.5% and 75% are almost tied, with the 75% mask ratio being just 0.015 percentage points higher. At 64k images, the 75% mask ratio becomes the best setting, with the 87.5% mask ratio becoming better than the 62.5% one. This pattern shows that smaller datasets benefit from a lower masking difficulty, while the original MAE mask ratio becomes stronger once more pre-training data is available.

The mask-ratio experiment was therefore used to define the setting for the final downstream experiment. For the smaller dataset sizes below 16k, the 62.5% mask ratio was used. For 16k images and larger, the original 75% mask ratio was used. This rule follows the main trend in Figure 3: smaller datasets benefit from an easier reconstruction task, while larger datasets can use the original MAE masking difficulty.

5.2 Experiment 2: How does downstream accuracy scale with dataset size?

Question. How does downstream CIFAR-10 accuracy change as the amount of MAE pre-training data increases?

Figure 4 shows the downstream CIFAR-10 test accuracy after MAE pre-training. The x-axis shows the number of Tiny ImageNet images used for pre-training. The y-axis shows the best CIFAR-10 top-1 test accuracy obtained by the downstream linear probe. The dashed line shows the random-weight baseline, where the same encoder is evaluated without MAE pre-training.

Downstream accuracy increases steadily as the amount of pre-training data grows. The model reaches about 59.6% CIFAR-10 accuracy after pre-training on 1k Tiny ImageNet images. Accuracy then rises to about 63.3% at 4k images, 67.0% at 16k images, and 69.5% at 100k images. The improvement becomes smaller after 32k images. This flattening should not necessarily be interpreted as full saturation with respect to data size, because the larger pre-training runs reached the 100k-step limit before satisfying the plateau criterion. The high-data settings may therefore still be undertrained compared with the smaller settings. All pre-trained models are far above the random-weight baseline, which shows that MAE learns useful transferable representations even with the smallest pre-training subset. The curve also shows that additional unlabeled data continues to improve the representation throughout the evaluated range.

Together, the two experiments answer the main empirical question of the paper. The mask-ratio experiment shows that the best masking difficulty is not fixed across dataset sizes. The downstream experiment shows that, under the selected mask-ratio rule, MAE representations improve as the amount of pre-training data increases.

6 Discussion

6.1 Interpretation of the results

The results show that MAE data efficiency depends on both dataset size and mask ratio. This is the main outcome of the experiments. MAE does not behave as if one masking setting is equally suitable for all data regimes. Instead, smaller datasets benefit from a lower mask ratio, while larger datasets can use the original 75% mask ratio more effectively. This matters because it shows that mask ratio should not be treated as a fixed MAE hyperparameter, but that it should be chosen with the available data budget in mind.

The mask-ratio experiment gives a possible explanation for this behavior. With a small pre-training dataset, each image has to provide more useful training signal. A lower mask ratio leaves more visible patches, so the model receives more context from each image during reconstruction. This can make the self-supervised task easier and more stable when there are

few images to learn from. A higher mask ratio removes more context, which may be useful at scale, but can make the task too difficult when the dataset is small. The improvement of the 62.5% mask ratio at 1k and 4k images supports this interpretation.

The result changes when more pre-training data is available. At 64k images, the 75% mask ratio gives the best validation accuracy. This suggests that once the dataset contains enough visual variation, the harder reconstruction task becomes useful again. In that setting, hiding more patches may encourage the encoder to learn broader image structure instead of relying on local cues. The conclusion is that lowering the mask ratio can improve data efficiency in the smallest data regimes, but the original MAE masking difficulty remains a strong choice once the dataset is larger.

The downstream CIFAR-10 results show that the selected MAE setup learns transferable representations across the evaluated dataset sizes. Accuracy increases as more Tiny ImageNet images are used for pre-training, and all pre-trained models are above the random-weight baseline. This means that MAE is useful even with limited pre-training data, but the benefit grows with additional unlabeled images. For resource-constrained training, the important point is not only that more data helps. The important point is that the training setup should also become easier when the dataset becomes smaller, otherwise the model may not use the limited data as effectively.

The stopping behavior gives an additional view of this data-efficiency trade-off. Runs with 16k images or fewer reached the plateau condition before the 100k-step limit. Runs above 16k images reached the 100k-step limit before plateauing. This suggests that smaller datasets stop providing new useful signal earlier, while larger datasets can still benefit from more gradient updates. In other words, larger datasets are not only better because they contain more images; they can also support longer useful training. This is important for resource-constrained training, because the best use of compute may depend on the dataset size. Small datasets may need a lower mask ratio and fewer updates, while larger datasets may need more updates before their full benefit is reached.

6.2 Limitations

The main limitation is that the largest dataset settings reached the 100k-step training limit before the monitored validation accuracy plateaued. The slower improvement after 32k images may therefore reflect the fixed update budget rather than a true saturation of MAE data efficiency. With a larger training budget, the 32k, 64k, and 100k settings might reach higher downstream accuracy, so the reported curve should be interpreted as data scaling under the chosen training protocol rather than as fully converged performance for each dataset size.

The mask-ratio experiment is also limited to three mask ratios and four dataset sizes. The results show a clear trend, but they do not identify the optimal mask ratio for every dataset size. For example, a value between 62.5% and 75% may work better near the transition around 16k images. A denser mask-ratio sweep would be needed to estimate the best masking schedule more precisely.

The experiments use one model size, one pre-training dataset, and one downstream dataset. All results are based on ViT-Tiny/8 pre-training on Tiny ImageNet and downstream evaluation on CIFAR-10. The same pattern may not hold for larger ViT models, higher-resolution images, or different downstream tasks. For this reason, the results should be read as evidence for this constrained setting, not as a general rule for all MAE training.

The final limitation is the checkpoint selection rule. Checkpoints were selected using Tiny ImageNet validation accuracy, while the final result is measured on CIFAR-10. This choice was made to avoid using the downstream task during pre-training. If CIFAR-10 accuracy were used for checkpoint selection, the training procedure could indirectly overfit to the downstream evaluation. Using Tiny ImageNet validation accuracy keeps the downstream task separate, but it also means that the selected checkpoint is not guaranteed to be the checkpoint with the best CIFAR-10 accuracy. Future work could test whether different task-independent checkpoint selection rules change the downstream scaling curve.

7 Responsible Research

This section reflects on the responsible research aspects of the study. The main goal of the project is to understand whether MAE can learn useful visual representations with less pre-training data. This goal is relevant beyond accuracy, because smaller data and compute requirements can make visual AI research more accessible. At the same time, reducing these requirements also raises ethical and environmental questions.

7.1 Ethical Considerations

This work studies data-efficient pre-training for vision models. A possible benefit is that smaller research groups can experiment with self-supervised visual learning without needing very large datasets or multi-GPU training. This can make research easier to reproduce and can reduce the gap between well-resourced and less-resourced institutions. In this sense, improving data efficiency can support more open and accessible visual AI research.

The same accessibility can also create risks. If useful visual models become easier to train, they may also become easier to use in harmful applications, such as unwanted surveillance or automated image analysis without consent. This paper does not develop such an application, but the method studied here could be used as part of one. For this reason, the results should be understood as a study of representation learning, not as a recommendation to deploy visual models without considering the application context.

The use of smaller datasets also has ethical limits. A smaller dataset can reduce compute cost, but it can also be less representative of the data seen in real applications. If a model is trained on a limited or biased subset, the learned representation may work worse for underrepresented groups or uncommon visual settings. This paper does not make fairness claims about the trained models. The results only show how MAE behaves under the tested dataset sizes and evaluation protocol.

7.2 Environmental Considerations

The experiments were designed to fit within a constrained compute budget. The model was small, the input resolution was limited to 64×64, and each run used a single GPU. These choices reduce the environmental cost compared with large-scale pre-training. The study also uses early stopping and checkpoint selection to avoid training longer than needed when the monitored representation no longer improves.

Despite those things, the total running cost was still non-negligible. Several MAE models were trained across multiple dataset sizes, mask ratios, and random seeds. The repeated runs were necessary to answer the research questions, but they increase the total compute used by the project. For this reason, the experimental design tries to balance scientific value with compute cost: the experiments are broad enough to study the main effects, but not expanded into a full hyperparameter search.

7.3 Reproducibility

The experimental setup was documented to make the results reproducible. Section 4 reports the datasets, subset construction, model configuration, training protocol, checkpoint selection rule, compute environment, and downstream evaluation protocol. These details define the full procedure needed to repeat the experiments.

The code used for the experiments is available at https://github.com/DDTerziev04/reducing-data-in-visual-ai. The repository includes the training scripts, configuration files, and dataset split definitions. Together, the reported setup and the released code should allow the experiments to be reproduced under the same conditions.

7.4 Use of Generative AI

Generative AI tools were used to support the writing process. They were used for drafting, restructuring, and improving clarity. The experimental design, implementation, results, and final claims were checked and edited by the author. The author remains responsible for the content of the paper.

8 Conclusion and Future Work

This paper studied how pre-training dataset size and mask ratio affect the data efficiency of Masked Autoencoders under constrained compute. The experiments show that MAE can learn useful transferable representations even from small unlabeled datasets. CIFAR-10 linear-probe accuracy increased as more Tiny ImageNet images were used for pre-training, which shows that additional unlabeled data improves the learned representation. At the same time, the mask-ratio experiment shows that the best masking difficulty depends on the amount of available pre-training data.

The main finding is that smaller datasets benefit from a lower mask ratio. For the 1k and 4k subsets, 62.5% masking gave better Tiny ImageNet validation accuracy than the original 75% MAE mask ratio. At 16k images, the two settings were almost equal. At 64k images, the original 75% mask ratio became the best setting. This suggests that small datasets need an easier reconstruction task, because each image must provide more useful training signal. Larger datasets can support a harder reconstruction task, because the model sees more visual variation and can still learn from less visible context.

These results answer the main research question by showing that MAE data efficiency is controlled by both dataset size and masking difficulty. The conclusion is that the mask ratio should not always be treated as a fixed default. When the amount of pre-training data is small, reducing the mask ratio can improve how effectively the model uses each image. When the dataset is larger, the original MAE masking difficulty becomes stronger again. This makes mask ratio an important part of designing data-efficient MAE training under limited compute.

Future work could first extend the experiments to larger datasets and longer training runs. In the current results, the 87.5% mask ratio becomes closer to the 75% mask ratio at larger dataset sizes. This suggests that the optimal mask ratio may continue to increase when even more pre-training data is available. A larger-scale study could test whether the best mask ratio eventually becomes higher than the original 75% setting. This would help determine whether the transition observed in this paper is part of a broader scaling pattern.

Future work could also study dynamic mask-ratio schedules instead of fixed mask ratios. One example is R2MAE, which samples the mask ratio during training rather than using one constant value [14]. Such a method could expose the model to both easier and harder reconstruction tasks during pre-training. Another example is curriculum masking, where the masking task changes during training. CL-MAE follows this idea by using a learnable masking module that gradually increases the complexity of the reconstruction task [15]. These approaches could be useful for data efficiency because the model would not have to rely on one fixed masking difficulty for the whole run.

Another direction is to replace random masking with adaptive patch selection. The original MAE masks patches uniformly at random, so it does not consider whether a patch contains important image content. Methods such as AutoMAE learn where to mask by using a mask generator that gives higher masking probability to informative patches while still controlling reconstruction difficulty [16]. Such approaches could be especially useful in small-data regimes, where each image has to provide as much useful learning signal as possible.

Finally, future work could test whether the same pattern holds for other models, datasets, and downstream tasks. This paper used ViT-Tiny/8, Tiny ImageNet, and CIFAR-10 to keep the experiments feasible under constrained compute. Larger Vision Transformers, higher-resolution datasets, and different downstream tasks may change the relation between dataset size and mask ratio. Testing these settings would show whether the findings are specific to this constrained setup or reflect a more general property of masked image modeling.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg,

A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[3] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119, 2020, pp. 1597–1607.

[5] S. Kornblith, J. Shlens, and Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2661–2671.

[6] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 10 347–10 357.

[7] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," in *International Conference on Learning Representations*, 2022.

[8] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9653–9663.

[9] J. H. Tan, "Pre-training of lightweight vision transformers on small datasets with minimally scaled images," 2024.

[10] Z. Xu, Y. Dai, F. Liu, W. Chen, Y. Liu, L. Shi, S. Liu, and Y. Zhou, "Swin MAE: Masked autoencoders for small datasets," *Computers in Biology and Medicine*, vol. 161, p. 107037, 2023.

[11] K. Wen, Z. Li, J. Wang, D. Hall, P. Liang, and T. Ma, "Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective," in *International Conference on Learning Representations*, 2025.

[12] A. Hägele, A. Dremov, A. Kosson, and M. Jaggi, "Scaling laws and compute-optimal training beyond fixed training durations," *arXiv preprint arXiv:2405.18392*, 2024.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 37, 2015, pp. 448–456.

[14] M. Dong, L. Wang, and Y. Kluger, "Understanding and enhancing mask-based pretraining towards universal representations," in *Advances in Neural Information Processing Systems*, 2025.

[15] N. Madan, N.-C. Ristea, K. Nasrollahi, T. B. Moeslund, and R. T. Ionescu, "Cl-mae: Curriculum-learned masked autoencoders," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2492–2502.

[16] H. Chen, W. Zhang, Y. Wang, and X. Yang, "Improving masked autoencoders by learning where to mask," in *Pattern Recognition and Computer Vision*, 2023.

A Experimental Configuration

Table 1: MAE pre-training configuration. The setup follows the original MAE recipe. The batch size is changed to 500 because larger batches did not fit in GPU memory and because 500 divides all subset sizes exactly. The mask ratio is varied to test how masking difficulty interacts with dataset size.

Setting	Value
Encoder	ViT-Tiny/8
Decoder hidden size	192
Decoder layers	4
Decoder attention heads	3
Decoder intermediate size	768
Pre-training data	Tiny ImageNet training subsets
Subset sizes	1k, 2k, 4k, 8k, 16k, 32k, 64k, 100k
Subset construction	Nested subsets
Objective	MAE pixel reconstruction
Optimizer	AdamW
Batch size (B)	500
Base learning rate (η_{base})	1.5×10^{-4}
Learning-rate scaling	$\eta_{\text{base}} \cdot B/256$
Weight decay	0.05
AdamW betas	$(0.9, 0.95)$
Mask ratios	62.5%, 75%, 87.5%
Augmentations	Random crop, random horizontal flip
Random seeds	0 and 42

Table 2: Downstream linear-probe configuration on CIFAR-10. The encoder is frozen, and only the probe head is trained. The reported metric is the best top-1 test accuracy across probe epochs.

Setting	Value
Downstream data	CIFAR-10
Evaluation type	Linear probing
Frozen model	ViT-Tiny/8 encoder
Pooling	Global average pooling
Feature normalization	$\mathrm{BatchNorm1d}(192)$, affine=false
Classifier	$\mathrm{Linear}(192, 10)$
Probe epochs	1000
Optimizer	SGD
Momentum	0.9
Weight decay	0
Base learning rate	0.1
Probe batch size	128
Effective learning rate	0.05
Learning-rate schedule	Cosine annealing to 0
Train transform	Resize to 64×64, ImageNet normalization
Test transform	Resize to 64×64, ImageNet normalization
Reported metric	Best top-1 test accuracy
Random seed	0