Reducing Data in Visual AI: Assessing Masked Autoencoder Data Efficiency Under Limited Compute

Dimo Terziev

EEMCS, Delft University of Technology, The Netherlands
Supervisors: Jan van Gemert, Petter Reijalt, Alex Manolache

1. Motivation and Research Question

Visual foundation models are often first trained on large image datasets and then reused for downstream tasks. In computer vision, this is commonly done with Vision Transformers, which split an image into small patches and process them with a transformer. However, training useful ViT representations often requires more data and compute than small research groups can access.

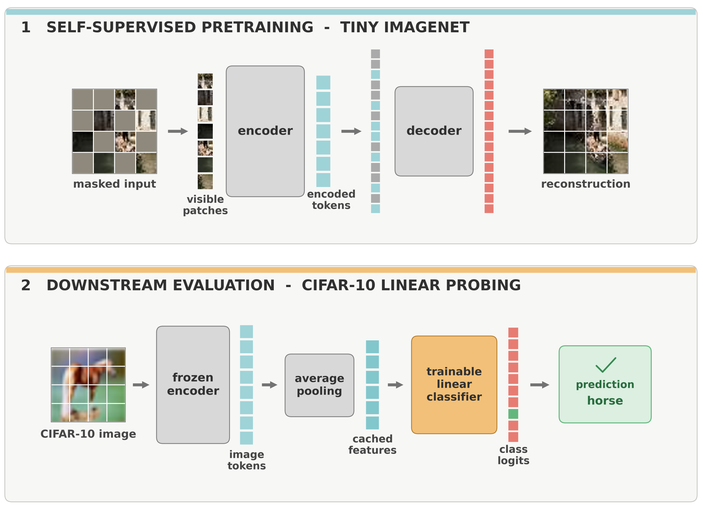
Study aim: Test how Masked Autoencoder (MAE) pre-training behaves when unlabeled data is limited. MAE learns without labels by hiding image patches and training the model to reconstruct them. The mask ratio controls how much of the image is hidden, and therefore how difficult the reconstruction task is.

Research question

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

- **Dataset-size question:** How does downstream task accuracy change as pre-training data grows from 1k to 100k images?
- **Masking question:** When pre-training data is limited, is the original 75% MAE mask ratio still the best choice?

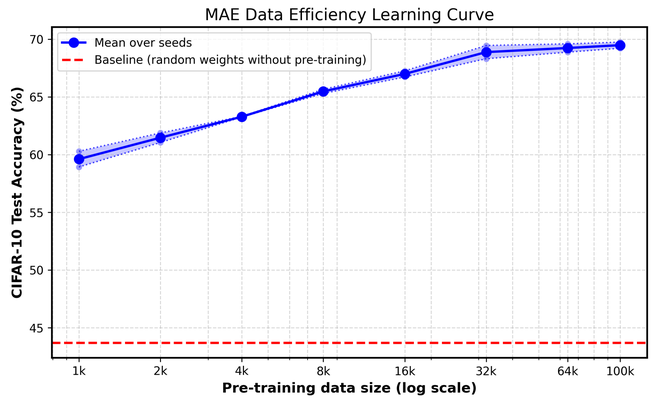
2. Experimental Design



Controlled pipeline: nested data subsets, MAE pre-training, frozen linear probing, and analysis.

Data: Tiny ImageNet subsets 1k–100k **Model:** ViT-Tiny/8 **Masks:** 62.5%/75%/87.5%
Schedule: Weight-Stable-Decay, 40 epoch warmup, 10% cooldown
Selection: Best plateau checkpoint, max 100k steps **Test:** CIFAR-10 linear probe

4. Main Result: More Unlabeled Data Helps Consistently

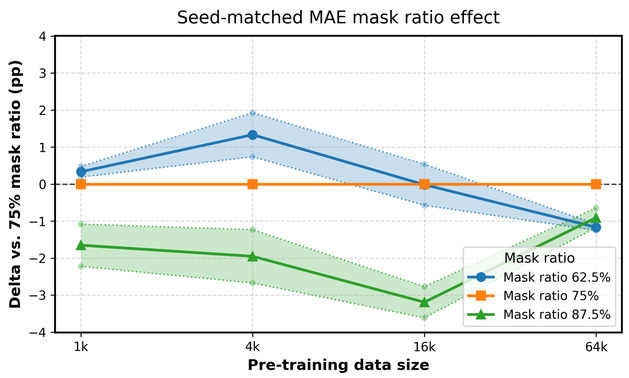


Accuracy rises from ∼59.6% at 1k to ∼69.5% at 100k; dashed red line is the random-weight baseline.

Finding 1 - MAE works at 1k.
Every pre-trained encoder is far above the random-weight baseline.

Finding 2 - Scaling continues.
Accuracy improves across 1k-100k; high-data gains may be capped by the 100k-step budget.

5. Mask Ratio Is a Data-Regime Hyperparameter



Difference relative to the original 75% MAE mask ratio at the same subset size.

- **1k-4k:** 62.5% masking is best; more visible context stabilizes learning.
- **16k:** 62.5% and 75% are almost tied.
- **64k:** 75% becomes strongest as harder reconstruction becomes useful.

6. Interpretation

Experiment 1: masking depends on data size.
Small subsets benefit from lower masking because more visible patches give each image a clearer signal. With more data, 75% masking becomes stronger because harder reconstruction becomes useful.

Experiment 2: more data improves transfer.
CIFAR-10 accuracy rises as Tiny ImageNet pre-training data grows from 1k to 100k images. MAE learns useful representations even from small unlabeled datasets, but continues to benefit from more data.

Caveat.
The largest runs hit the 100k-step limit before plateauing, so slower improvement after 32k should not be read as full saturation.

Takeaway: MAE data efficiency depends on both dataset size and masking difficulty: lower masking helps small data, while 75% becomes stronger with more data.

7. Limitations and Future Work

- **Constrained setting:** ViT-Tiny/8, Tiny ImageNet, and CIFAR-10; larger models or tasks may shift the threshold.
- **Training cap:** 32k-100k runs hit 100k steps before plateauing, so high-data results may be undertrained.
- **Sparse mask sweep:** Only 62.5%, 75%, and 87.5% were tested; the transition region needs denser sweeps.

Future work: Longer high-data runs, dynamic or curriculum masking, adaptive patch selection, larger ViTs, and more downstream tasks.

8. Responsible Research, Contact, and References

- **Accessibility:** Data-efficient pre-training can lower barriers for smaller research groups.
- **Caution:** Easier training does not remove risks such as biased or unwanted visual AI use.
- **Reproducibility:** Splits, scripts, configs, and code are documented.

Dimo Terziev – EEMCS, TU Delft
D.D.Terziev-1@student.tudelft.nl
Code: github.com/DDTerziev04/reducing-data-in-visual-ai

Refs: Bommasani et al. 2021; Dosovitskiy et al. 2021; He et al. 2022; Caron et al. 2021.

AI disclosure: generative AI supported text restructuring and LaTeX formatting; the author retains responsibility for final claims.