Replicating Masked Autoencoders: Scalable Self-Supervised Learning on Tiny ImageNet

Leo Burgund

February 2026

Note: All the code used for this project is available at `https://github.com/Leob000/m2-MAE-reproduction/`.

Abstract

This report investigates the reproducibility and optimization dynamics of Masked Autoencoders (MAE) as described by He et al. [5]. By implementing a ViT-Lite architecture on the Tiny ImageNet dataset, we explore how generative pre-training through high-ratio patch masking serves as a powerful optimization proxy for visual representation learning.

1 Motivations

The recent explosion in the capacity of deep learning architectures has shifted the primary bottleneck of computer vision from model design to data acquisition. Modern models easily overfit even large-scale datasets like ImageNet-1K, creating an insatiable demand for labeled data that relies heavily on expensive and slow human annotation.

In Natural Language Processing (NLP), this bottleneck was effectively bypassed through self-supervised learning (SSL) methods such as BERT [3] and GPT [7]. These models utilize masked or autoregressive objectives to learn from vast amounts of unlabelled text. However, adapting these techniques to computer vision has historically faced two major hurdles:

- **Architectural Disparity:** Until recently, Convolutional Neural Networks (CNNs) were the dominant architecture. Convolutions operate on regular grids, making it difficult to integrate ”mask tokens” or handle missing data without breaking the spatial structure. The advent of Vision Transformers (ViT) [4] solved this, as they treat images as sequences of independent patches, allowing for seamless masking.
- **Information Density:** Images and language differ fundamentally in their information density. Language is a human-generated signal that is highly semantic and dense; masking even a few words creates a challenging task. In contrast, images are natural signals with heavy spatial redundancy. A missing pixel can often be recovered by simple interpolation from its neighbors without any high-level understanding.

To address this, MAE introduces a specific strategy: a very high masking ratio (e.g., 75%). This shifts the optimization objective from low-level pixel interpolation to a ”hard” pretext task that requires a holistic semantic reconstruction of the scene. By forcing the model to infer large missing regions, the optimization process is coupled with the learning of high-level visual concepts, effectively mirroring the success of masked language modeling in the visual domain.

2 Model Architecture

The MAE framework is a specialized autoencoder built upon the Vision Transformer (ViT) [4] architecture. It masks a high proportion of image patches and tasks the model with reconstructing the missing content in pixel space. Its core innovation lies in an asymmetric encoder-decoder design optimized for both computational efficiency and powerful representation learning.

Before detailing the MAE framework, it is necessary to briefly review the standard Vision Transformer (ViT) [4]. Unlike CNNs, which process pixels via local receptive fields, a ViT treats an image as a sequence of tokens.

An input image $x \in \mathbb{R}^{H \times W \times C}$ is divided into a regular grid of non-overlapping patches of resolution $P \times P$. These patches are flattened and mapped via a learned linear projection to a D-dimensional latent space, yielding a sequence of $N = \frac{HW}{P^2}$ patch embeddings. Because the transformer architecture is inherently permutation-invariant, positional embeddings are added to each patch token to retain spatial information.

The entire sequence is then processed by a stack of standard Transformer blocks, consisting of alternating layers of Multi-Head Self-Attention (MHSA) and Multi-Layer Perceptrons (MLP). The critical optimization bottleneck in this architecture is the self-attention mechanism, which scales quadratically with the sequence length, i.e., $\mathcal{O}(N^2)$. As N grows large for high-resolution images, training becomes prohibitively expensive.

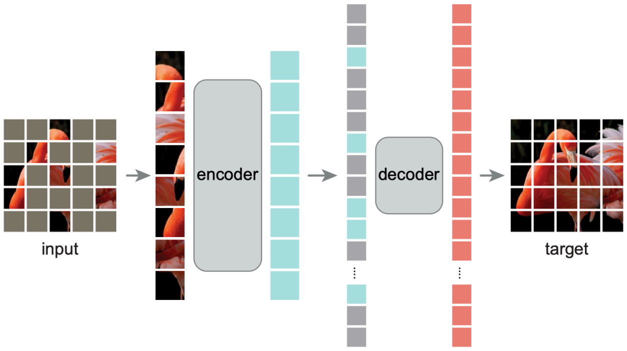

Figure 1: MAE architecture overview. Figure from [5].

2.1 Asymmetric Encoder-Decoder

The key to MAE's efficiency is its asymmetric structure, which treats visible and masked patches differently.

Encoder The encoder is a standard ViT that processes **only the visible patches** (e.g., 25% of the input). This simple design choice is a profound hardware-conscious optimization. By drastically reducing the sequence length of tokens fed into the computationally heavy Transformer blocks, it cuts down the quadratic complexity of the self-attention mechanism. This enables MAE to scale to much larger models within a reasonable compute budget, a central goal of the paper.

Decoder The decoder is a lightweight, temporary ViT that is discarded after pre-training. Its role is to reconstruct the full image from two inputs: 1) the encoded representations of the visible patches from the encoder, and 2) a series of learnable "mask tokens" that serve as placeholders for the missing patches. Critically, the full sequence of patches is only processed by this shallow, narrow decoder, which performs less than 10% of the overall computation. This design makes the decoder's architecture independent of the encoder, allowing for flexible configurations.

2.2 Reconstruction Objective

The model's optimization target is the reconstruction of the original image pixels.

Loss Function The loss is a simple Mean Squared Error (MSE) computed between the decoder's output and the original pixel values of the **masked patches only**. This serves as a direct, effective proxy for learning visual representations without requiring complex mechanisms like contrastive pairs or external tokenizers used in other SSL methods.

Target Normalization Following the original paper, we normalize the pixel values of each patch on a per-patch basis before computing the loss. This local normalization enhances contrast and was found to improve representation quality during fine-tuning.

3 Optimization Dynamics of the Masked Objective

From an optimization perspective, MAE reformulates self-supervised learning by shifting from the dominant paradigm of contrastive learning to a direct generative objective. Contrastive methods, such as SimCLR [2] or MoCo [6], learn representations by maximizing the agreement between differently augmented views of the same image (positive pairs) while simultaneously minimizing the agreement with views from other images (negative pairs). This objective explicitly encourages the model to learn features that are linearly separable. However, it often relies heavily on careful data augmentation and large batch sizes to prevent representation collapse. MAE bypasses these requirements.

Let M denote the set of indices corresponding to the masked patches, and let x_i and $\hat{x}_i$ represent the normalized ground-truth pixels and the predicted pixels for the i-th patch, respectively. The optimization objective is defined strictly over the masked regions:

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

By purposefully excluding the visible patches from the loss computation, the objective avoids a trivial local minimum where the autoencoder learns an identity mapping. Instead, the high masking ratio forces the network to project the sparse visual context into a lower-dimensional manifold, learning robust, non-linear, and semantic representations capable of hallucinating the missing 75% of the signal.

4 Experimental Reproduction on Tiny ImageNet

To validate the scalability and generalizability of MAE under constrained compute budgets, we replicate the pre-training and evaluation protocol on the Tiny ImageNet (TIN) dataset (100,000 images, 200 classes, 64 × 64 resolution). The original paper uses ImageNet-1K.

4.1 Setup and Compute Constraints

Following the work of Charisoudis et al. [1], who demonstrated a successful MAE reproduction on a smaller scale, we utilize a ViT-Lite backbone tailored for smaller resolutions. The encoder consists of 7 Transformer blocks with an embedding dimension of 256, while the decoder is strictly lightweight, consisting of only 2 blocks with a dimension of 128. This results in an encoder size of roughly 3.7M parameters. This deliberate downscaling ensures that the full pre-training pipeline can be executed efficiently on a single Nvidia A100 GPU within a few hours.

Full details regarding the learning rate schedules, weight decay, and augmentation strategies for all phases (Pre-training, Linear Probing, and Fine-Tuning) are documented in Appendix A.

4.2 Results

We evaluate the quality of the learned representations through two standard protocols: Linear Probing and end-to-end Fine-Tuning.

Our pre-training phase involved training the ViT-Lite model for 1200 epochs, saving a checkpoint every 100 epochs. As shown by the training loss curve in Figure 2, the pre-training objective converged successfully, with the loss continuously decreasing even towards the end of the 1200 epochs, which is consistent with the original MAE paper [5].

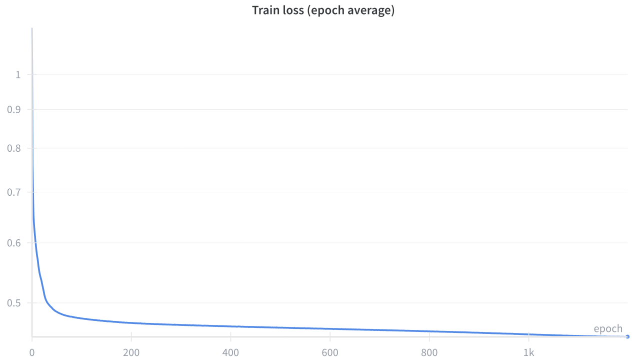

Figure 2: Pre-training loss (MSE on masked patches), averaged per epoch.

For downstream evaluation, we performed both linear probing and fine-tuning using various checkpoints saved during pre-training. For linear probing, we froze the pre-trained encoder weights and trained a new linear classification head on top of its features. For fine-tuning, we unfroze the entire model (encoder and a randomly initialized classification head) and trained it end-to-end. The primary evaluation, shown in Figure 3, compares the downstream performance of these two evaluation methods.

Our best fine-tuned model achieved a Top-1 accuracy of approximately **62.2%** on the Tiny ImageNet validation set, derived from a model pre-trained for 700 epochs. In contrast, the best linear probing result was **23.78%**, obtained from a model pre-trained for only 100 epochs.

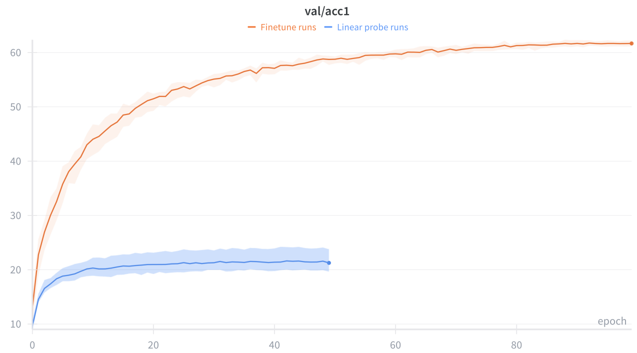

Figure 3: Top-1 accuracy on Tiny ImageNet validation set for Fine-tuning vs. Linear Probing. We trained each method using different pre-training checkpoints (from epoch 100 to 1200) and grouped the results by the evaluation method.

Our experiments reveal a significant 38-point gap between the linear probing and fine-tuning accuracies. While the original MAE paper [5] also noted a discrepancy, the effect is magnified in our scaled-down ViT-Lite architecture. This suggests that while the masked reconstruction objective effectively captures rich visual semantics, these features are highly entangled and non-linear. Contrastive methods, as previously discussed, explicitly optimize for linear separability and thus tend to yield better linear probing results. However, MAE's objective prioritizes strong, non-linear feature maps that, when allowed to adapt via fine-tuning, demonstrate excellent transfer learning capabilities despite the reduced model capacity. The poor linear probing performance may not be an indication of poor features, but rather that the features are not organized in a linearly separable way.

It is interesting to note that while the pre-training loss continued to decrease even after 1000+ epochs, the fine-tuning performance remained relatively stable across different pre-training checkpoints (e.g., fine-tuning from epoch 700 yielded similar results to later checkpoints). Linear probing, however, showed more variance depending on the pre-training checkpoint. This suggests that beyond a certain point, further reductions in reconstruction loss during pre-training may not significantly improve the quality of features for downstream fine-tuning tasks on this specific dataset, but could still subtly influence the linear separability of features.

4.3 Qualitative Analysis and Generalization

To provide a more intuitive understanding of the model's capabilities, we include a supplementary Jupyter Notebook that offers a hands-on implementation of the core MAE mechanics, including patchifying, masking, and reconstruction (see Figure 4).

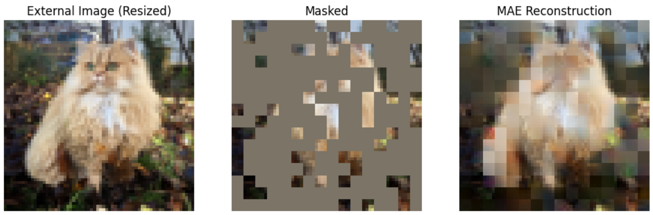

Figure 4: Reconstruction of an out-of-distribution image. The model captures the core semantic essence of the subject despite a 75% mask.

5 Discussion and Limitations

The scalability offered by the reduced encoder size (offered by the masking) and the lightweight temporary decoder is a massive advantage, allowing for the training of data-hungry architectures like ViT-Huge. Furthermore, unlike contrastive learning paradigms (e.g., MoCo, SimCLR) that heavily depend on carefully engineered data augmentations to prevent representation collapse, MAE achieves strong representations with minimal augmentation (e.g., random cropping), relying instead on the difficulty of the masking task itself for regularization, hence allowing a training with less risks of overfitting.

However, the approach is not without limitations. First, the optimization objective relies purely on pixel-level Mean Squared Error (MSE), which encourages exact color and texture reproduction rather than strictly high-level semantic understanding.

Second, as hinted in our Tiny ImageNet experiments, MAE features lack the immediate linear separability seen in contrastive methods. While the objective forms strong non-linear representations, extracting maximum value from these features requires end-to-end fine-tuning or tuning an MLP head, which is computationally more expensive downstream than simple linear probing.

6 Conclusion

In this report, we investigated the optimization dynamics and scalability of Masked Autoencoders under strict resource constraints. By pre-training a ViT-Lite architecture on the Tiny ImageNet dataset, we validated that a high-ratio masking objective serves as a highly efficient generative proxy for visual representation learning. Crucially, our evaluation revealed a defining characteristic of MAE optimization: while the generative objective captures rich, high-level semantics that yield strong end-to-end fine-tuning performance (62.2% Top-1 accuracy), the resulting feature space seems highly non-linear. This explains the significant performance gap when using linear probing, contrasting sharply with the linearly separable features prioritized by contrastive learning paradigms. Ultimately, this reproduction confirms that while generative pre-training provides profound computational advantages during the encoding phase, the choice of downstream adaptation is critical to unlocking its full representational power.

References

[1] Athanasios Charisoudis, Simon Huth, and Emil Jansson. [re] masked autoencoders are small scale vision learners. *ReScience C*, 9(2), 2023.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

A Appendix: Hyperparameter Configurations

Table 1 details the configurations used across the three phases of our reproduction, adapting the original MAE hyperparameters to the ViT-Lite architecture on Tiny ImageNet.

Table 1: Hyperparameters for ViT-Lite on Tiny ImageNet.

Parameter	**Pre-training**	**Fine-tuning**	**Linear Probing**
Epochs	1200	100	50
Batch Size	256	128	128
Base LR	5e-4	1e-3	1e-3
Weight Decay	0.15	0.05	0.0
Warmup Epochs	20	5	2
Optimizer	AdamW	AdamW	AdamW
Momentum	(0.9, 0.95)	(0.9, 0.999)	(0.9, 0.95)
Mask Ratio	75%	—	—
Pixel Norm	Yes	—	—
Augmentation	Crop + Flip	RandAug(2, 9)	Flip