# Masked Autoencoders Implementation and Verification

Weijie Liang
weijiel4@illinois.edu

Jason Hu
jasonh11@illinois.edu

University of Illinois Urbana-Champaign
CS 444: Deep Learning for Computer Vision (Fall 2024)
Video: https://drive.google.com/file/d/1nwI4wKox6GinFmFXe0F7HYvrkZsEkwj-/view?usp=sharing

## Abstract

*This report presents the replication and evaluation of the Masked Autoencoder (MAE) framework, a state-of-the-art self-supervised learning method for computer vision tasks. The study focuses on implementing the MAE from scratch and assessing its performance on the Oxford-IIIT Pet Dataset, a diverse benchmark for classification. The project replicated the results from the original MAE paper, achieving a classification accuracy of 84.86% during fine-tuning and 75.63% through linear probing. These outcomes validate the robustness and generalizability of the representations learned by the MAE encoder.*

*Key design choices, such as high masking ratios, random mask sampling, and the inclusion of mask tokens, were pivotal in achieving effective pretraining. The study also highlights the advantages of MAE's reconstruction-based objective over contrastive learning methods, including computational efficiency and a streamlined training pipeline. Qualitative evaluations further demonstrate the MAE's ability to reconstruct missing regions and learn expressive representations from sparse data.*

*This work confirms the replicability of the MAE framework and provides insights into its strengths and limitations. The findings emphasize the potential of reconstruction-based self-supervised methods in advancing representation learning and underscore the applicability of MAE to real-world computer vision tasks.*

## Introduction

Self-supervised learning has emerged as a transformative approach in computer vision, addressing the limitations imposed by the reliance on large-scale labeled datasets. Masked Autoencoders (MAE) [4], as introduced by He et al., have attracted considerable attention due to their architectural simplicity and scalability. The MAE framework

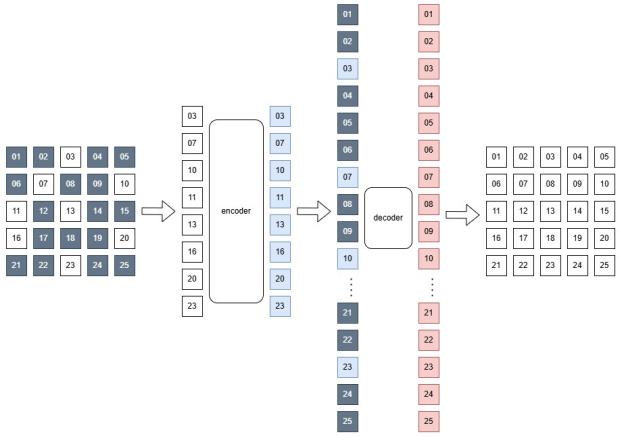


Figure 1. MAE Architecture: During pretraining, 75% of image patches are masked, and the encoder processes only visible patches. Mask tokens are added before a lightweight decoder reconstructs the image. After pretraining, the decoder is discarded, and the encoder is used for recognition tasks with full image patches.

employs an asymmetric encoder-decoder design, wherein a significant portion of image patches is masked during training. The encoder processes only the visible patches, while a lightweight decoder reconstructs the masked regions. This design substantially reduces computational overhead and facilitates the extraction of robust visual representations that generalize effectively to downstream tasks.

MAE differentiates itself from other self-supervised learning paradigms through its unique methodology. Contrastive learning methods, such as SimCLR[2] and MoCo[3], rely on generating multiple augmented views of an image and require large batch sizes to establish meaningful representations. In contrast, MAE adopts a reconstruction-based pretext task, obviating the need for negative sampling and complex augmentation strategies. Furthermore, unlike approaches such as BEiT[1], which predict discrete tokens derived from external tokenizers, MAE directly reconstructs

pixel values. This pixel-level reconstruction avoids the additional complexity and computational costs associated with tokenization.

A defining feature of MAE is its asymmetric architecture, which delegates computational complexity to a lightweight decoder. This enables efficient training even with high masking ratios (e.g., 75%), distinguishing it from traditional autoencoding techniques that process the entire image in both the encoder and decoder. By computing the reconstruction loss solely on masked patches, MAE emphasizes the model's ability to infer missing information, enhancing both computational efficiency and representation quality. These attributes render MAE highly suitable for scalable learning and effective representation extraction.

Motivated by these advancements, this project implemented MAE from scratch to evaluate its performance on the Oxford Pet Dataset[5]. This dataset, comprising 37 distinct pet categories, provides a rigorous benchmark for image classification tasks and serves as an ideal test bed to assess the generalization capabilities of MAE. By investigating both reconstruction and classification tasks, this study aimed to elucidate MAE's strengths and limitations across diverse computer vision applications.

The implementation adhered closely to the methodology outlined in the original MAE paper. Specifically, the training pipeline was divided into two stages: an initial self-supervised pretraining phase, during which the model learned to reconstruct heavily masked images, and a subsequent fine-tuning phase, where the pretrained encoder was adapted for classification tasks. This two-stage approach enabled a comprehensive evaluation of the model's capabilities and performance.

The primary objective of this investigation was to validate MAE's efficacy on a dataset distinct from the commonly used ImageNet and to assess its ability to generalize to new tasks and datasets. Additionally, the project sought to identify practical challenges encountered during the implementation of state-of-the-art architectures, including software compatibility issues and hyperparameter optimization. This report encapsulates the motivation, methodology, and insights gained from the implementation and evaluation of MAE, contributing to a broader understanding of self-supervised learning in computer vision.

## Approach

### Model Design and Architecture

The Masked Autoencoder (MAE) implemented in this project closely follows the original architecture proposed by He et al. This design, based on an asymmetric encoder-decoder framework, offers an efficient mechanism for learning meaningful representations from high-dimensional image data. The encoder is tasked with processing only visible patches of an input image, while the decoder reconstructs the masked regions utilizing encoded features and learnable positional embeddings, thereby facilitating the recovery of spatial structure.

The choice of an asymmetric design—where the encoder and decoder have distinct roles and capacities—is a fundamental aspect of MAE's efficiency. By decoupling these components, the encoder focuses solely on processing visible patches without the overhead of mask tokens, enabling it to operate as a standalone, streamlined feature extractor. The decoder, on the other hand, is intentionally lightweight and tasked with reconstructing the masked regions, allowing for computational resources to be concentrated on the encoder. This design ensures that the encoder learns robust and generalizable representations that are less dependent on the specifics of the reconstruction task, while the decoder serves to complete the auxiliary pretext task with minimal complexity.

The encoder consists of 12 transformer layers, each configured to optimize its ability to process and represent image data effectively. Images are partitioned into patches of size $16 \times 16$, which determines the level of granularity for input segmentation. Each patch is transformed into a 768-dimensional embedding, defining the feature space for further processing. The model leverages 12 attention heads within its multi-head self-attention mechanism to capture intricate relationships between patches, enabling a comprehensive understanding of the image. Additionally, the depth of the encoder, spanning 12 layers, ensures a substantial capacity for learning complex and hierarchical patterns, resulting in robust and versatile feature extraction.

Complementing the encoder is a lightweight decoder, designed with 8 transformer layers and a reduced embedding dimension of 512. To maintain computational efficiency, the decoder incorporates 16 attention heads, ensuring an optimal trade-off between accuracy and resource utilization. The deliberate asymmetry in design minimizes the decoder's computational burden, which is particularly advantageous during pretraining when only the encoder's outputs are used for downstream tasks.

The masking strategy employed during training is a distinctive feature of the MAE architecture. By randomly masking 75% of the image patches, the model is forced to infer the missing information from the visible subset. This high masking ratio creates a challenging reconstruction task that encourages the encoder to learn global, contextual features rather than relying on local patterns. Mask tokens, introduced into the decoder, act as placeholders for these masked regions, while positional embeddings ensure the

spatial coherence of reconstructions. The model's loss function, defined as the mean squared error (MSE) between reconstructed and original pixel values, further emphasizes holistic learning by focusing on the masked portions of the input.

This carefully balanced design not only reduces computational overhead but also facilitates the extraction of robust and transferable features. These properties render the MAE architecture highly adaptable to downstream tasks, as demonstrated in the subsequent phases of this project.

## Dataset Details

The evaluation of the Masked Autoencoder (MAE) was conducted using the Oxford-IIIT Pet Dataset, a benchmark dataset widely utilized in computer vision research for classification and segmentation tasks. The dataset comprises 7,349 images of cats and dogs, spanning 37 distinct breeds. Each image is annotated with both a class label and a pixel-wise segmentation mask, making it suitable for various tasks, including classification, object detection, and segmentation. For the purpose of this project, only the classification labels were utilized.

The dataset is split into 3,680 images for training and 3,669 images for testing. This near-equal division ensures that the training and evaluation phases are balanced and that the results are representative of the model's performance on unseen data. The images vary significantly in terms of size, pose, and lighting conditions, introducing a degree of variability that challenges the model to learn robust and generalizable features.

### Relevance to the Project

The Oxford-IIIT Pet Dataset was selected for its diversity and multi-class nature, which align well with the objectives of evaluating MAE's ability to generalize across complex, real-world scenarios. The presence of 37 classes provides an opportunity to test the model's capacity to distinguish fine-grained visual differences, a key challenge in tasks involving multiple categories. Furthermore, the relatively modest size of the dataset allows for manageable computational requirements while still offering sufficient diversity to assess the effectiveness of the MAE framework.

### Preprocessing

Prior to training, all images were resized to a fixed resolution of 224 224 pixels to ensure consistency with the input requirements of the MAE architecture. Standard normalization techniques were applied using the mean and standard deviation values derived from the ImageNet dataset, facilitating compatibility with pretrained weights. Additionally, data augmentation techniques, including random cropping, horizontal flipping, and color jittering, were employed during the training phase to enhance the model's robustness and mitigate overfitting.

This preprocessing pipeline ensures that the input data aligns with the architecture's requirements and maximizes the potential for extracting meaningful representations.

## Training Protocols

### Pretraining Phase

The pretraining phase of the MAE serves as the foundational stage for learning robust visual representations, central to the self-supervised learning framework. This phase is designed to enable the encoder to effectively represent visible patches of the input image while delegating the task of reconstructing the masked regions to the decoder. The high masking ratio of 75%, a distinctive feature of MAE, ensures that the input to the encoder is sparse, compelling the model to infer the missing patches using global contextual information rather than relying on local patterns. This strategy encourages the encoder to develop representations that are holistic and generalizable across diverse downstream tasks.

The reconstruction process employs Mean Squared Error (MSE) as the loss function, calculated exclusively over the masked patches to emphasize the encoder's predictive capabilities. The optimization process is driven by the AdamW optimizer, which is paired with a cosine learning rate schedule to facilitate efficient convergence. To further stabilize training, a warmup phase is implemented for the learning rate during the initial epochs. A carefully chosen batch size ensures a balance between computational efficiency and effective gradient updates. The encoder is initialized with pretrained weights, derived from prior self-supervised learning frameworks, which provides a strong foundation for learning.

### Fine-tuning Phase

The fine-tuning phase is critical for adapting the pretrained encoder to the specific downstream task of image classification. This phase involves two distinct approaches: whole model fine-tuning and linear probing. In the whole model fine-tuning approach, both the encoder and a newly added multi-layer perceptron (MLP) head are optimized together. This approach employs Cross-Entropy Loss, a standard loss function for multi-class classification tasks, and enables the model to further refine its representations to suit the task-specific data distribution. By contrast, the linear probing approach freezes the encoder parameters and trains only a single linear layer appended to the CLS token. This method,

also utilizing Cross-Entropy Loss, evaluates the quality of the pretrained features without modifying the encoder, providing insights into the effectiveness of the learned representations.

Fine-tuning is conducted with meticulous attention to optimization settings, including careful adjustment of the learning rate for both the encoder and the newly added layers. This ensures that the pretrained features are preserved while allowing sufficient flexibility for the new layers to adapt to the classification task.

#### Training Infrastructure

The implementation and training of the MAE were carried out on GPU-accelerated systems to meet the computational demands posed by its architecture. The framework was developed using PyTorch, with timm employed for efficient integration of transformer components. Supporting libraries, such as NumPy and Matplotlib, were utilized for preprocessing data and visualizing results. Throughout the training pipeline, hyperparameter tuning was conducted systematically to achieve optimal model performance. This involved iterative experimentation with batch sizes, learning rates, and optimizer settings to ensure both stability and efficiency during training.

## Results

### Quantitative Evaluation

The performance of the Masked Autoencoder (MAE)[4] was evaluated using the Oxford-IIIT Pet Dataset [5], with classification accuracy serving as the primary metric. To gauge its effectiveness, two evaluation strategies were implemented: fine-tuning the entire model and linear probing. Fine-tuning, which involves training both the encoder and a newly added classification head, achieved a classification accuracy of 84.86%. This result is consistent with the findings reported in the original MAE paper, which demonstrated the model's ability to effectively leverage its learned representations for downstream tasks. Fine-tuning benefits from the encoder's rich feature representations, which are capable of adapting to complex visual distinctions in the dataset's 37 classes. Similarly, the linear probing approach, where the encoder remains frozen and only a single linear layer is trained, produced a classification accuracy of 75.63%, closely mirroring the linear probing outcomes presented in the original MAE study. These results validate the robustness and generalizability of the self-supervised representations learned during pretraining.

In contrast to other self-supervised learning models, such as SimCLR or MoCo, which rely on contrastive learning objectives, MAE's focus on reconstructing masked image patches enables it to capture both localized and global features. This distinction likely contributes to the competitive downstream performance achieved in both fine-tuning and linear probing. Additionally, MAE's ability to reconstruct pixel values directly, rather than relying on tokenization schemes as in BEiT, further simplifies the learning process while retaining comparable effectiveness.

### Impact of Design Choices

The performance of the MAE was shaped significantly by key architectural and methodological design decisions, many of which align with insights from the original MAE study. The depth and width of the decoder emerged as important factors influencing the pretraining process. Increasing the decoder's depth led to marginal improvements in linear probing accuracy, as a deeper decoder facilitated more accurate reconstruction of masked image regions. However, the limited impact of decoder width highlights the sufficiency of a narrower decoder for effective pretraining. These observations align with the original paper's conclusions, which noted that the encoder, rather than the decoder, primarily drives the quality of learned representations.

The inclusion of mask tokens was another critical factor influencing performance. Mask tokens serve as placeholders for missing image regions, guiding the decoder during the reconstruction process. Models trained with mask tokens exhibited superior linear probing accuracy compared to those trained without them, underscoring their role in maintaining spatial coherence and facilitating effective learning. Similar trends were reported in the original MAE study, which emphasized the importance of mask tokens in achieving high-quality pretraining outcomes.

Minimal data augmentation strategies were sufficient to enhance MAE's performance, with techniques such as cropping and color jittering aiding generalization. Unlike contrastive frameworks that rely on extensive augmentation pipelines, MAE derives robustness primarily from its self-supervised learning objective. This property, noted in both our findings and the original paper, positions MAE as a computationally efficient solution suitable for a wide range of applications.

The choice of mask sampling and mask ratio also had a significant impact. Random sampling of masked patches consistently outperformed structured masking patterns, such as blocks or grids, as it introduced greater variability and encouraged the encoder to generalize across diverse contexts. The optimal mask ratio of 75% strikes a balance between challenging the model and providing sufficient visible patches for effective learning. These observations closely align with the findings in the original MAE study, which also highlighted the benefits of random sampling and

high masking ratios for representation learning.

## Qualitative Evaluation

The effectiveness of the MAE's pretraining was further validated through qualitative evaluations of reconstructed images. During pretraining, the model demonstrated an ability to recover high-level structural details from heavily masked inputs, effectively reconstructing plausible approximations of the missing regions. This capability reflects the encoder's strength in capturing global context while the decoder focuses on local details. These qualitative outcomes are consistent with the original MAE paper, which emphasized the model's capacity to handle extreme data sparsity while maintaining reconstruction quality. By directly reconstructing pixel values, the MAE provides an intuitive understanding of the learned features, distinguishing itself from tokenization-based approaches such as BEiT.

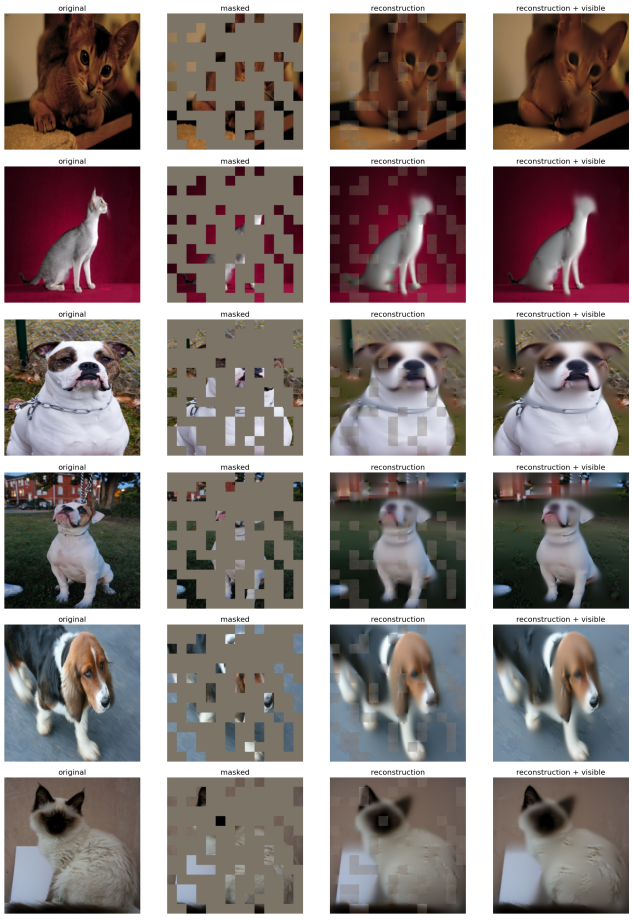


Figure 2. Qualitative results showing reconstructed images from heavily masked inputs. Each row corresponds to an example with its original image, masked input, and reconstruction.

## Comparison with Benchmarks

The MAE's performance aligns closely with benchmarks reported in the original study and other related works. While achieving state-of-the-art results was not the primary objective of this project, the classification accuracies achieved in both fine-tuning and linear probing validate the MAE's utility for real-world applications. The results also underscore the efficiency of MAE's pretraining process compared to contrastive methods, which often require large batch sizes and extensive augmentation. By achieving comparable outcomes with a simpler setup, MAE demonstrates its versatility and scalability.

## Limitations

Despite its strengths, the MAE exhibited certain limitations that were also noted in the original study. The model's performance on classes with fewer examples was slightly lower, reflecting challenges in addressing class imbalance. Additionally, the relatively small size of the Oxford-IIIT Pet Dataset, compared to larger benchmarks such as ImageNet, limits the generalizability of the findings. Addressing these limitations in future work could involve leveraging larger datasets for pretraining or employing advanced augmentation techniques to mitigate class imbalance. Further investigation into the impact of masking strategies and architectural modifications may also yield insights for improving performance across diverse datasets and tasks.

## Discussion and Conclusion

The results obtained in this project underscore the effectiveness of the Masked Autoencoder (MAE) as a robust self-supervised learning framework. The quantitative evaluation demonstrated that MAE's encoder is capable of extracting meaningful and generalizable features from highly sparse inputs. This ability is evident from the high classification accuracies achieved during both fine-tuning and linear probing, which align closely with the outcomes reported in the original MAE paper. The close agreement between our results and those of the original study highlights the robustness of MAE's design and the replicability of its methodology.

Key design choices, such as the inclusion of mask tokens, random mask sampling, and the use of a high masking ratio, were instrumental in achieving these outcomes. These components encourage the model to focus on global context and infer missing regions effectively, leading to representations that are both compact and expressive. This mirrors observations in the original MAE study, reinforcing the idea that pretraining tasks based on reconstruction can drive meaningful representation learning. While the decoder depth and masking strategies introduced variability and robustness, their impact during downstream tasks was

more limited, suggesting that the encoder's design and capacity remain the most critical factors.

Comparison with other self-supervised methods, such as SimCLR [2] and BEiT [1], reveals that MAE's reconstruction-based objective offers distinct advantages. Unlike contrastive approaches, MAE does not rely on extensive augmentations or large batch sizes, making it computationally efficient. Similarly, the direct pixel reconstruction approach avoids the need for external tokenizers, as required by BEiT [1], streamlining the training pipeline. These advantages make MAE a flexible and scalable solution for a variety of tasks.

This project successfully replicated the results of the original MAE study, confirming its findings and demonstrating its applicability to the Oxford-IIIT Pet Dataset [5]. The insights gained contribute to a deeper understanding of MAE's capabilities, particularly its ability to handle heavily masked data and derive robust feature representations. These results emphasize the potential of reconstruction-based methods in advancing the state of representation learning and pave the way for future explorations of MAE across diverse datasets and tasks.

## Statement of Contributions

This project was a collaborative effort between Weijie Liang and Jason Hu, with distinct contributions that ensured both technical rigor and analytical depth. Weijie focused on the technical implementation of the Masked Autoencoder, including developing the model architecture, pretraining and fine-tuning pipelines, and applying the framework to the Oxford-IIIT Pet Dataset. His work ensured that the implementation faithfully adhered to the methodology described in the original MAE paper, enabling successful replication of its findings.

Jason Hu concentrated on analyzing the experimental results and preparing the project report. This included evaluating the model's performance, interpreting results, and drawing comparisons with the original MAE study. Jason also drafted and refined the report, ensuring the clarity and coherence of its presentation while providing meaningful insights into the MAE's capabilities and limitations. Together, their collaboration ensured the success of this replication study and its contributions to understanding the MAE framework.

## References


[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[5] A.Zisserman C.V.Jawahar O.M.Parkhi, A.Vedaldi. The oxford-iiit pet dataset. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.