Rethinking Patch Dependence for Masked Autoencoders

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we present cross-attention masked autoencoders (CrossMAE). This framework employs only cross-attention in the decoder to independently read out reconstructions for a small subset of masked patches from encoder outputs, yet it achieves comparable or superior performance to traditional MAE across models ranging from ViT-S to ViT-H. CrossMAE challenges the necessity of interaction between mask tokens for effective masked pretraining and leads to much more efficient pretraining. Code is available here.

1 Introduction

Masked image modeling [20, 13, 25, 3] is a crucial unsupervised learning technique in computer vision. A notable approach is masked autoencoders (MAE), where the model reconstructs missing pixels from a small subset of visible image patches. MAE efficiently pre-trains large models on vast datasets, yielding strong results on various tasks [14, 16, 21].

MAE uses *self-attention* across both visible and masked tokens. However, masked tokens lack information, and it's unclear whether self-attention between masked tokens is beneficial. We break down the decoding process into self-attention among masked tokens and cross-attention with visible tokens. Notably, cross-attention dominates, with a score of 1.42 versus 0.39 for self-attention. This raises two questions: **1)** Is self-attention between masked tokens necessary? **2)** Can masked patches be reconstructed independently from the encoder, speeding up pretraining?

To address these questions, we propose CrossMAE, which differs from MAE in three key ways:

1. **Cross-attention for decoding:** Mask tokens act as queries in a cross-attention decoder to reconstruct patches from visible tokens, reducing sequence length and computation.
2. **Independent partial reconstruction:** Mask tokens are decoded independently based on visible token features, allowing for faster pretraining by decoding fewer patches.
3. **Inter-block attention:** CrossMAE uses features from different encoder blocks, leveraging both low- and high-level features for better learning.

Our results show that CrossMAE reconstructs images coherently without masked token interactions, relying on the encoder's global context. Performance remains strong, proving the encoder captures necessary information for reconstruction. **Our main contributions are:**

1. **We offer a new understanding of MAE.** Our findings reveal that MAE reconstructs coherent images through the encoder's global representation, not through interactions between masked patches in the decoder. The model performs well even without such interactions, showing the encoder effectively captures the necessary global information.
2. **We propose replacing self-attention with a cross-attention readout.** Since the MAE encoder captures the full global representation, we recommend replacing self-attention in the decoder with

Submitted to NeurIPS 2023 Workshop: Self-Supervised Learning - Theory and Practice. Do not distribute.

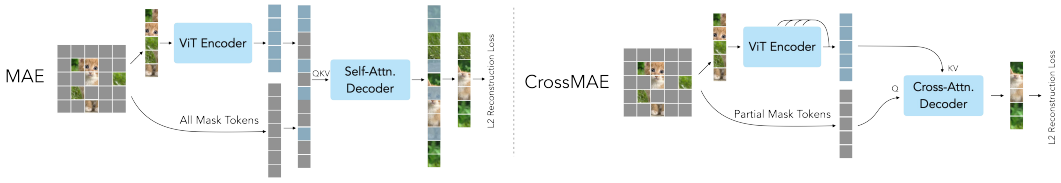

Figure 1: MAE [13] concatenates *all* mask tokens with the visible patch features from a ViT encoder and passes them to a decoder with self-attention blocks to reconstruct the original image. Patches that correspond to visible tokens are then dropped, and an L2 loss is applied to the rest of the reconstruction as the pretraining objective. CrossMAE instead uses cross-attention blocks in the decoder to reconstruct only a subset of the masked tokens.

cross-attention to independently aggregate encoder outputs for each input token, eliminating the need for token-to-token communication in the decoder.

3. **CrossMAE offers similar or better performance with lower computational costs** for tasks like image classification and instance segmentation compared to MAE, across vision transformer models from ViT-S to ViT-H.

2 Related Works

Self-Supervised Learning. In self-supervised learning, models are trained on tasks where the supervision comes from the input data itself, without labels. Contrastive learning methods, like SimCLR [6], CPC [19], and MoCo [12], learn by contrasting positive and negative samples. Other methods, such as BYOL [11], iBOT [26], and DINO [4], train a student model to imitate a teacher model without using negative pairs.

Masked Modeling. Masked modeling learns representations by reconstructing missing parts of the input. In natural language processing (NLP), early works like BERT [8] and its extensions [18, 15] introduced masked language modeling to enable few-shot learning with bidirectional transformers. In computer vision, early studies like Stacked Denoising Autoencoders [24] and Context Encoder [20] used masked image modeling for denoising and representation learning. With the rise of transformers in vision tasks [9], researchers have explored adapting language sequence modeling to vision transformers. BEiT [2], MAE [13], and SimMIM [25] applied BERT-style pretraining to vision transformers. Unlike NLP models, MAE and SimMIM found that much higher mask ratios are needed for effective visual representation learning. Recent works have expanded masked pretraining to hierarchical architectures [25, 17] and explored the role of data augmentation [5, 10].

3 CrossMAE

3.1 Preliminaries: Masked Autoencoders

Masked Autoencoders (MAE) [13] pretrain Vision Transformers (ViTs) [9] by dividing input images into patches, selecting a random subset as visible. The ViT encoder processes the visible patches and a learnable `[CLS]` token to produce a set of feature latents. These, with masked patch positional embeddings and a learnable mask token, are input to the MAE decoder. Both the encoder and decoder use transformer blocks with self-attention. The decoder's output length matches the original input and assumes visible-masked token interactions. A final fully connected layer reconstructs the image, with loss applied only to masked positions. We explore simplifying the decoding process by eliminating self-attention among masked tokens, while maintaining the model's downstream performance.

3.2 Reconstruction with Cross-Attention

We replace the self-attention in the decoder with cross-attention, using it to decode the encoder's latent embeddings into pixel values. The decoder uses multi-head cross-attention, where queries come from previous decoder blocks (or from the masked patch positions for the first block), while keys and values are from the encoder features. We use weighted means of encoder feature maps (in Section 3.4 to serve as the keys and values for the decoder layers. Residual connections refine the decoded tokens as they pass through decoder blocks. Unlike the original Transformer [23], CrossMAE's decoder skips the self-attention layer, allowing tokens to be decoded independently. We find that removing

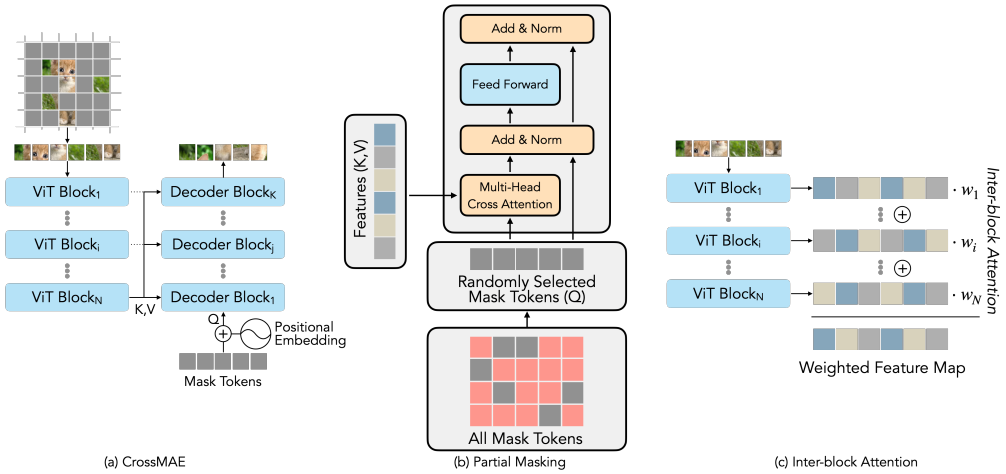

Figure 2: Overview of CrossMAE. (a) The vanilla version of CrossMAE uses the output of the last encoder block as the keys and queries for cross-attention. The first decoder block takes the sum of mask tokens and their corresponding positional embeddings as queries, and subsequent layers use the output of the previous decoder block as queries to reconstruct the masked patches. **(b)** Unlike the decoder block in [23], the cross-attention decoder block does not contain self-attention, decoupling the generation of different masked patches. **(c)** CrossMAE's decoder blocks can leverage low-level features for reconstruction via inter-block attention. It weighs the intermediate feature maps, and the weighted sum of feature maps is used as the key and value for each decoder block.

the self-attention layer does not lead to a reduction in downstream performance and improve training efficiency (see the appendix for more detail). While MAE projects encoder features to the decoder space using an MLP, CrossMAE performs the projection in the multi-head cross-attention module. CrossMAE does not limit itself to a single cross-attention block but stacks multiple decoder blocks, similar to the traditional Transformer [23].

3.3 Partial Reconstruction

CrossMAE's decoder uses cross-attention instead of self-attention, allowing independent decoding of mask tokens. This enables partial reconstruction of specific spatial locations, unlike MAE which requires all masked tokens due to self-attention. We introduce a "prediction ratio" (γ) as the ratio of predicted tokens to all image tokens, with $\gamma \in (0, p]$ where p is the mask ratio. Reconstructing only a random subset of masked locations maintains the expected mean square error loss but increases variance by (p/γ). Adjusting the learning rate to $\gamma\beta/p$ compensates for this. This partial reconstruction approach reduces computational complexity while preserving representation quality, as cross-attention scales linearly with the number of masked tokens.

3.4 Inter-block Attention

MAE's self-attention decoder creates an information bottleneck by using only the last encoder block's features with mask tokens. CrossMAE's cross-attention decoder allows different decoder blocks to access features from various encoder blocks. To select which encoder features to use, we propose learnable inter-block attention for feature fusion. This approach takes a weighted sum of visible token embeddings across encoder blocks at the same spatial location, combining multi-block features for each decoder block. Specifically, each decoder block uses a weighted linear combination of encoder feature maps as keys and values. For a token t_k in decoder block k of a model with encoder depth n, we initialize weights $w^k \in \mathcal{R}^n \sim \mathcal{N}(0, 1/n)$. Then, $t_k = \sum_{j=1}^{n} w_j^k f_j$, where f_j are encoder feature maps. This method of using weighted features from different encoder blocks significantly improves CrossMAE's performance compared to using only the last block's features.

4 Experiments

We perform self-supervised pretraining on ImageNet-1K, following MAE [13]'s hyperparameter settings, only modifying the learning rate and decoder depth. The hyperparameters were initially

Method	ViT-S	ViT-B	ViT-L	ViT-H
Supervised [22]	79.0	82.3	82.6	83.1
DINO [4]	-	82.8	-	-
MoCo v3 [7]	81.4	83.2	84.1	-
BEiT [2]	-	83.2	85.2	-
MultiMAE [1]	-	83.3	-	-
MixedAE [5]	-	83.5	-	-
CIM [10]	**81.6**	83.3	-	-
MAE [13]	78.9	83.3	**85.4**	85.8
CrossMAE (25%)	79.2	83.5	**85.4**	86.3
CrossMAE (75%)	79.3	**83.7**	**85.4**	**86.4**

Table 1: *ImageNet-1K classification accuracy*. CrossMAE performs on par or better than MAE. All experiments are run with 800 epochs. The best results are in **bold** while the second best results are underlined.

Method	AP^{box}		AP^{mask}	
	ViT-B	ViT-L	ViT-B	ViT-L
Supervised [16]	47.6	49.6	42.4	43.8
MoCo v3 [7]	47.9	49.3	42.7	44.0
BEiT [2]	49.8	53.3	44.4	47.1
MixedAE [5]	50.3	-	43.5	-
MAE [16]	51.2	54.6	45.5	48.6
CrossMAE	**52.1**	**54.9**	**46.3**	**48.8**

Table 2: *COCO instance segmentation.* Compared to previous masked visual pretraining works, CrossMAE performs favorably on object detection and instance segmentation tasks.

determined on ViT-Base and then directly applied to ViT-Small, ViT-Large, and ViT-Huge. Both CrossMAE and MAE are trained for 800 epochs.

4.1 ImageNet Classification

Setup. The model performance is evaluated with end-to-end fine-tuning, with top-1 accuracy used for comparison. We compare two versions of CrossMAE: one with a prediction ratio of 25% (1/3 of the mask tokens) and another with 75% (all mask tokens). Both models are trained with a mask ratio of 75% and a decoder depth of 12.

Results. As shown in Table 1, CrossMAE outperforms vanilla MAE using the same ViT-B encoder in terms of fine-tuning accuracy. This shows that replacing the self-attention with cross-attention *does not degrade* the downstream classification performance of the pre-trained model. Moreover, CrossMAE outperforms other self-supervised and masked image modeling baselines, *e.g.*, DINO [4], MoCo v3 [7], BEiT [2], and MultiMAE [1]. We also conduct ablations in Appendix A.

4.2 Object Detection and Instance Segmentation

Setup. We additionally evaluate models pretrained with CrossMAE for object detection and instance segmentation, which require deeper spatial understanding than ImageNet classification. Specifically, we follow ViTDet [16], a method that leverages a Vision Transformer backbone for object detection and instance segmentation. We report box AP for object detection and mask AP for instance segmentation, following MAE [13].

Results. As listed in Table 2, CrossMAE, with the default 75% prediction ratio, performs better compared to these baselines, including vanilla MAE. This suggests that similar to MAE, CrossMAE performance on ImageNet positively correlates with instance segmentation. Additionally, CrossMAE's downstream performance scales similarly to MAE as the model capacity increases from ViT-B to ViT-L. This observation also supports our hypothesis that partial reconstruction is suprisingly sufficient for learning dense visual representation.

4.3 Training Throughput and Memory Utilization

Due to partial reconstruction and confining attention to between mask tokens and visible tokens, CrossMAE improves pre-training efficiency over MAE. According to our experimental results, the FLOPs reduction does translate to an $1.54\times$ training throughput and at least 50% reduction in GPU memory utilization compared to MAE.

5 Discussion and Conclusion

Our study shows that image reconstruction in MAE is driven by the encoder's global representation, not patch interactions in the decoder. We simplify the decoder by using cross-attention to aggregate encoder outputs, tested on models from ViT-S to ViT-H. This approach matches or outperforms traditional methods in image classification and instance segmentation while being more efficient. CrossMAE is scalable and well-suited for large-scale visual pretraining, especially with video data.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022.

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

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022.

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[5] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22742–22751, 2023.

[6] 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.

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019.

[9] 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. In *ICLR*, 2020.

[10] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. In *The Eleventh International Conference on Learning Representations*, 2023.

[11] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[12] 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.

[13] 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 (CVPR)*, pages 16000–16009, 2022.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[15] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

[16] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision*, pages 280–296. Springer, 2022.

[17] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. *arXiv:2205.13137*, 2022.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[19] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[20] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[21] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023.

[22] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

[24] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

[25] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022.

[26] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

Method	Acc. (%)
MAE	83.0
CrossMAE	**83.3**
CrossMAE + Self-Attn	83.3

(a) Attention type in decoder blocks. Adding back self-attention between mask tokens does not improve performance.

Mask Ratio	Acc. (%)
65%	**83.5**
75%	83.3
85%	83.3

(b) Mask ratio. CrossMAE has consistent performance across high mask ratios.

Pred. Ratio	Acc. (%)
15%	83.1
25%	83.2
75%	**83.3**

(c) Prediction ratio. CrossMAE performs well even when only a fraction of mask tokens are reconstructed.

# Feature Maps Fused	Acc. (%)
1	82.9
3	83.3
6	**83.5**
12	83.3

(d) Inter-block attention. A combination of six select encoder feature maps is best.

Decoder Depth	Acc. (%)
1	83.0
4	83.1
8	83.1
12	**83.3**

(e) Decoder depth. CrossMAE performance scales with decoder depth.

Image Resolution	Acc. (%)
224	83.2
448	**84.6**

(f) Input resolution. CrossMAE scales to longer input sequences.

Table 3: *Ablations on CrossMAE*. We report fine-tuning performance on ImageNet-1K classification with 400 epochs (*i.e.*, half of the full experiments) with ViT-B/16. MAE performance is reproduced using the official MAE code. Underline indicates the default setting for CrossMAE. **Bold** indicates the best hyperparameter among the tested ones. 1 feature map fused (row 1, Table 3(d)) indicates using only the feature from the last encoder block. We use 25% prediction ratio for both settings in Table 3(f) to accelerate training.

A Ablations

In our ablation studies, we found that CrossMAE, with its cross-attention decoder, outperforms vanilla MAE in downstream tasks, as shown in Table 3a, and combining cross-attention with self-attention does not improve performance, indicating cross-attention alone is sufficient. CrossMAE also effectively learns representations by reconstructing as few as 15% of tokens, compared to the 100% required by vanilla MAE, with minimal impact on downstream fine-tuning performance, as shown in Table 3b and Table 3c. Additionally, as detailed in Table 3d, using multiple encoder feature maps in the inter-block attention mechanism enhances performance, with the best results from using six feature maps. While a deeper 12-block decoder slightly improves performance, CrossMAE achieves similar results with just one block, as shown in Table 3e, demonstrating its efficiency. Furthermore, we found that models with lower prediction ratios benefit more from deeper decoders, as seen in **??**. Lastly, increasing input resolution improves classification accuracy, indicating that CrossMAE can scale well with longer input sequences, as observed in Table 3f.