A Mix-and-Mask Approach to Self-Supervised Image Pretraining

Christos Polzak
Department of Computer Science
Stanford University
clcp@stanford.edu

Joy Yun
Department of Computer Science
Stanford University
joyyun@stanford.edu

Sam Xing
Department of Computer Science
Stanford University
samxing@stanford.edu

Abstract

In recent years, deep vision models have shown remarkable performance on standard benchmarks, yet they continue to struggle with specific error types, such as those related to texture and object occlusion. The Architecture-Agnostic Masked Image Modeling (A2MIM)[10] approach has demonstrated promising improvements in the robustness and generalization of self-supervised convolutional neural networks (CNNs), particularly in enhancing resilience to occlusion. This paper examines how modifications to the mask patch size. and subsequently the layer at which masking is applied, of the A2MIM framework impacts model robustness to various error factors, highlighted by ImageNet-X. We show that modifying the mask patch size has minor but ambiguous implications for a model's texture and occlusion robustness, and that reconstruction quality can be misaligned with robustness. Through this project, we aim to provide a deeper understanding of how these modifications affect model performance and robustness, contributing valuable insights into the development of more resilient deep vision models.

1. Introduction

In the search for deeper, more complex, and more robust vision models, self-supervised learning (SSL) has emerged as a key paradigm, allowing for models to be trained on vastly more data without requiring human annotation, and indeed, several of the currently highest-performing vision models, especially those intended for transfer to a variety of downstream tasks, involve a substantive SSL pre-training step [7, 2, 11].

One promising approach to self-supervised learning is Masked Auto-Encoding (MAE) [7]. MAE draws from the popular Vision Transformer (ViT) architecture [5], which maps image patches of a fixed size (commonly 16x16) into vector embeddings that are then passed through a Transformer model [14] that is essentially identical to those used in the natural language processing (NLP) domain. Indeed, MAE itself, also referred to as masked image modeling (MIM), is an adaptation of the masked language modeling (MLM) self-supervised task frequently used for NLP Transformer models, and it works by masking out a subset of the image patch embeddings and tasking the model to fill the masked patches back in, while focusing on middle-order interactions among patches.

However, as vision models continue to grow in popularity and sophistication, understanding their limitations and explicitly improving upon them is crucial to ensuring the safety and reliability of their applications. To this end, a few efforts have been made to analyze and benchmark common sources of computer vision model failures. Some of these efforts include ImageNet-A and ImageNet-O, benchmark datasets adversarially-designed by humans to test models trained on the ImageNet dataset, and ImageNet-X, which selects prototypical images for each class and directly annotates how examples in the ImageNet-1k validation set vary relative to the prototype. These datasets demonstrate that evaluating the robustness of a model solely based on ImageNet accuracies is insufficient. Texture and occlusion, in particular, are common failure points among models.

Extensive research has demonstrated that MAE turns out to be a very robust pretraining task, and the original MAE paper resulted in tremendous improvements on many challenge datasets, including ImageNet-A. Additionally, A2MIM has challenged the notion that MAE only works for Transformer-based architectures, and demonstrated through extensive experiments that their proposed A2MIM framework works effectively with both CNNs and Transformers, while improving performance on various benchmarks.

However, even using MAE, the ImageNet-X paper showed that handling occlusion is still a major challenge for models [10]. In this work, we conduct systematic experiments to explore how changes to the mask patch size used, and subsequently the layer at which the masking is applied at, in the A2MIM system can increase models' robustness to the consistently challenging issues of occlusion and texture variation. The input into our model is an image. Random patches of the image are masked out and inputted into a CNN (ResNet-50) pretrained on ImageNet-100 using our modified A2MIM to generate a reconstructed image.

2. Related Works

Existing work proposes other methods for pre-training deep vision networks with a focus on generalization and robustness. In addition to the MAE and MIM methods mentioned above, MoCoV3 [3] minimizes the contrastive loss to distinguish augmented versions of an image, DINO [2, 11] utilizes a self-supervised teacher-student framework, and CLIP [13] requires a model to match images and their captions. The pretraining methods have the shared goal of the model learning high-quality visual representations for images.

Another common pretraining method is masked predictions, which has been leveraged in both NLP and CV. In NLP, BERT [4] is trained on Masked Language Modeling (MLM) which requires the model to classify randomly masked input tokens. In CV, CNNs are trained on self-supervised inpainting [12] and colorization tasks [15], which require the model to perform context-based pixel prediction and colorize a grayscale image, respectively. Similarly, ViTs [5] are trained to reconstruct missing patches of an image masked out by Masked Auto Encoders (MAE) [6] and BEiT (Bert pre-training of image transformers) [1]. While most existing work in masked image modeling (MIM) applies to ViTs, A2MIM extends modern MIM training to CNNs [10], achieving comparable performance.

3. Methods

3.1. Overview

By integrating MIM with CNNs, A2MIM offers significantly greater flexibility to adjust the image patch size during pretraining. Inspired by the observation from the ImageNet-X paper that models struggle with texture and occlusion and leveraging the patch size flexibility provided by A2MIM, in this paper we investigate how variations in the pretraining task, especially regarding patch size and mask ratio, affect the downstream robustness of the model.

3.2. Model Architecture and Task

Building off of the A2MIM setup, we use ResNet-50 with an image reconstruction decoder during pretraining. When masking the images, we replace the masked regions with the mean pixel value of the image to alleviate local statistic distortions, which reduces noise and allows the model to focus on modeling more informative medium-frequencies in these regions.

In the A2MIM implementation, during training, learnable mask tokens are placed at layer 0 in the ResNet-50 architecture where the receptive fields were masked. The masking operation for MIM is shown below:

$$x_{mask} = x \odot (1 - M) + T \odot M \tag{1}$$

where x is the input image, M is the random occlusion mask, and T is the learnable mask token.

The A2MIM paper posits that masking at the stem layer, or too early in the model, undermines the CNN's context extraction abilities. Masking at input layers of the model also distracts it from learning necessary low-level feature extractions. Therefore, we follow A2MIM's suggestion to mask intermediate features at a point when the feature representations contain both semantic and spatial information.

Due to the spatial properties of CNNs and the changing size of the feature representations throughout the network, the layer at which the masking is applied is intrinsically tied to the mask patch sizes that can be explored, and vice versa. Given that there is 2x2 pooling occurring at each layer, the pixel at location (0,0) in the original feature representation does not directly map to the pixel at (0,0) after the next layer's transformations. We need to ensure that the parts of the image we mask end up directly mapping to distinct features in the layer at which we add the mask token. In order to mask at a patch size of NxN, we need to ensure that N is divisible by the receptive field of the layer at which the masking is applied. The masking operation is shown below:

$$z^l_{mask} = z^l + T \odot D(M) \tag{2}$$

where z^l is the current feature map at stage-l in the CNN and $D(\odot)$ is the corresponding down-sampling function for the occlusion mask.

3.3. Loss Function

For our loss function we used the Fourier loss from A2MIM that allows us to encourage a model to learn middle-order interactions by placing weights on different ordered frequency interactions. More specifically, A2MIM adds on a loss L_{freq} in Fourier domain to the traditional spatial domain loss L_{spa}, such that total loss becomes:

$$\mathcal{L} = \mathcal{L}_{spa} + \lambda \mathcal{L}_{freq} \tag{3}$$

The $\mathcal{L}_{freq}$ loss uses a dynamic weighting matrix to steer the model to focus on learning specific ordered frequency interactions in an image at different points in the training. The dynamic weights matrix prioritizes learning on-the-fly hard frequencies or, in other words, where there is the greatest absolute difference between the model's reconstructed frequencies and the raw frequencies. While this most often leans towards larger weights for middle-level frequencies because they are the hardest to learn, we unfortunately cannot explicitly modify the weights matrix always place greater weight on a particular frequency.

3.4. Data

In this project, we pretrain and finetune our models using the ImageNet-100 dataset. ImageNet-100 is a subset of the ImageNet-1K dataset that consists of 130,000 images for training, 5,000 images for validation, and 10,000 images for testing. These images are categorized into 100 different classes, each representing a distinct object category. Note that we used ImageNet-100 over ImageNet-1K due to time and compute constraints.

Before pretraining on ImageNet-100, we perform several preprocessing operations on the dataset. We begin by resizing each image to 224x224. Next, each image is normalized and standardized, and then each image is randomly cropped and resized. Finally, a random horizontal flip is performed. These preprocessing steps introduce variability in the training data, which helps in making the model more robust and less likely to overfit.

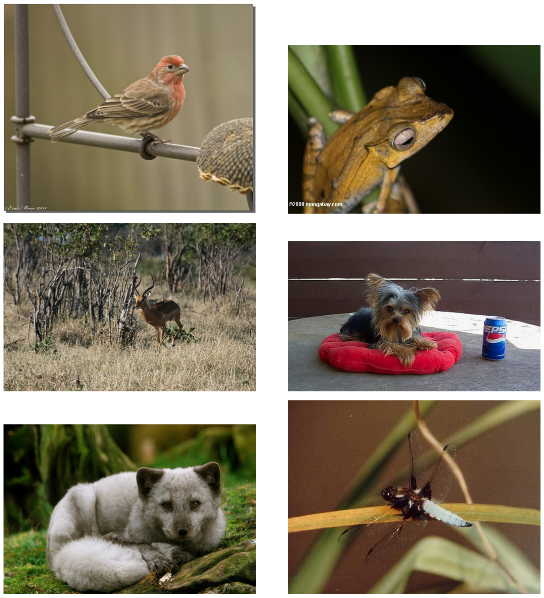

Figure 1. Example images from the ImageNet-100 dataset. From left to right and top to bottom, the labels are: House Finch, Tree Frog, Impala, Yorkshire Terrier, Arctic Fox, and Dragonfly.

3.5. Training Hyperparameters

We ran pretraining for 20 epochs using AdamW with a learning rate of 1e-3 and weight decay of 0.05 for all experiments. Finetuning ran for 15 epochs, also using AdamW and with the same learning rate of 1e-3 and weight decay of 0.05 for all experiments; standard cross-entropy was used for the classification.

3.6. Evaluation Metrics

ImageNet-X [9], introduced by Idrissi et al. in 2022, is an extension of the ImageNet dataset that is designed to provide a detailed understanding of the weaknesses of deep vision models. ImageNet-X selects prototypical images for each class and directly annotates how (all) examples in the ImageNet-1k validation set vary relative to the prototype (e.g. different background, lighting, object pose, texture). Consequently, models evaluated via ImageNet-X can now have their errors distinguished by the different variation factors. In this project, we utilize ImageNet-X to evaluate how our modifications to the A2MIM algorithm impact the errors identified in ImageNet-X, particularly those related to texture and occlusion.

4. Experiments and Results

4.1. Baselines

First, we corroborate the occlusion robustness claim set forth by the A2MIM paper [10] by comparing ResNet-50s [8] pre-trained on the full ImageNet-1k, and then evaluated on the full 1000 classes of ImageNet-X. Specifically, we compare A2MIM against MoCoV3, another popular self-supervised learning method. As shown in Figure **??**, the A2MIM model is generally less susceptible to errors related to occlusion, with substantially lower error ratios on cases of object or person blocking and similar performance on objects in partial view. We also verify the ImageNet-X paper's claim that vision models struggle with both texture and occlusion, as both models have high error ratios on the related ImageNet-X factors (for texture: ‘texture‘ and to some extent ‘subcategory‘; for occlusion, ‘person blocking‘, ‘object blocking‘, and ‘partial view‘).

4.2. Variations in Patch Size and Masking Ratio

Next, we begin modifying the A2MIM training pipeline to study the effects of patch size and mask ratio on the specific occlusion and texture-related ImageNet-X factors. Note that masked patches are square and that the patch size refers to the one-dimensional contiguous length of the image masked by a single patch; meanwhile, mask ratio refers to the total percentage of the image (rounded up to the nearest patch size) that is masked out (larger value means more is masked out).

Training from scratch on ImageNet-100, we investigate several combinations of MIM pre-training, then evaluate the models after finetuning them on the labeled version of ImageNet-100:

1. Patch Size = 4, Mask Ratio = 0.6 and 0.8
2. Patch Size = 32, Mask Ratio = 0.6 and 0.8
3. Patch Size = 56, Mask Ratio = 0.4 and 0.6

The results of these experiments are reported in Table 1. Succinctly, we make several observations:

1. Our masking hyperparameters have a minimal effect on overall validation accuracy on ImageNet-100, regardless of what patch size is used.
2. There is a consistent correlation between increasing the masking ratio and obtaining robustness to texture; the effect of mask ratio is far greater than the effect of the patch size itself.
3. Smaller patch sizes result in improved robustness to objects in partial view.
4. Patch size and mask ratio seem uncorrelated with a model's downstream ability to detect subcategory. However, a patch size of 4 with a mask ratio of 80% is the standalone best model in subcategory robustness.

4.3. Mixed Masking

Intrigued by our mixed-bag of results for different patch sizes and mask ratios, we investigate the potential for variations in the masking pipeline throughout training. Concretely, we modify the A2MIM pipeline such that every batch of training examples is randomly masked in one of two different ways, then let pretraining converge, followed by finetuning on the labeled ImageNet-100. Specifically, we test the following combinations:

1. Patch Size = 4, Mask Ratio = 0.6 OR Patch Size = 56, Mask Ratio = 0.4
2. Patch Size = 4, Mask Ratio = 0.6 OR Patch Size = 32, Mask Ratio = 0.6

These results are also reported in Table 1. These experiments seem fairly inconclusive: both mixed-mask models surpass their single-strategy counterparts with regard to robustness to partial viewing of the classified object; however, the two mixed models trade off in being worse at dealing with variations in either texture or subcategory, respectively, and do not have overall better validation performance.

5. Discussion

Initially, we hypothesized that since smaller patch sizes necessarily force the model to fill in lower-level details, a smaller patch size would result in greater understanding and consequently robustness to texture; by contrast, we expected that larger patch sizes would push the model to make greater inferences about occluded objects, as substantial portions of the objects could be masked out. Unfortunately, our quantitative results fail to substantiate this claim: ImageNet-100 lacks sufficient examples in the blocking categories of ImageNet-X to properly analyze occlusion, and texture robustness seems more a function of mask ratio than patch size.

However, qualitative analysis reveals our hypothesis to be half-right but slightly misaligned. When looking at the actual examples of reconstruction, as shown in Figure 3, it is readily apparent that the smaller patch size results in a higher-fidelity reconstruction of the original image, even when substantially more of the original image is masked out. This seems to imply that we are correct in saying that smaller patch size allows the model to focus more on the high-fidelity details of the image. The strange part is that although the image reconstructions are better, the actual robustness to texture hardly changes, suggesting that understanding reconstruction of texture and robustness to texture variation during classification are actually quite misaligned. Furthermore, despite the substantially poorer ability to reconstruct images, both models result in ultimately similar validation accuracy, highlighting either a flaw in our methodology or that perhaps MIM and masked image reconstruction are not necessarily the best strategies for reliable and robust pre-training. Undoubtedly, there is further work to be done.

6. Future Work

Given additional time and computational resources, future work would extend our investigation into several promising areas. One key avenue would be to train using ImageNet-1K instead of ImageNet-100. Doing so could provide more comprehensive insights into the scalability and robustness of our modified A2MIM approach. Due to short deadlines, extensive training times, and lack of computational resources, we were not able utilize ImageNet-1K. Additionally, integrating adversarial robustness evaluations and real-world deployment scenarios would offer a more holistic view of the models' resilience. These expanded studies would deepen our understanding of masked image modeling and its potential to enhance the robustness and generalization of deep vision models.

Table 1. Comparison of model overall and factor accuracies (%) on ImageNet-X

Model	Background	Color	Partial View	Pattern	Pose	Shape	Subcategory	Texture	Val Acc
Patch=4, Ratio=60	67.5	57.7	44.2	66.3	76.8	27.3	46.2	38.1	27.3
Patch=4, Ratio=80	69.1	61.2	66.7	68.3	77.3	40.9	34.6	52.4	34.6
Patch=56, Ratio=40	69.9	63.2	68.3	69.2	78.9	45.4	50.0	42.8	42.9
Patch=56, Ratio=60	69.8	60.5	73.0	69.2	78.1	40.9	50.0	52.4	40.9
Patch=32, Ratio=60	70.9	59.6	73.0	70.1	79.7	36.4	50.0	47.6	36.4
Patch=32, Ratio=80	68.1	60.5	71.4	66.3	78.3	45.4	53.8	52.3	45.4
Patch=4/56, Ratio=60/40	69.3	61.2	76.1	67.7	78.5	45.5	38.5	52.4	38.5
Mixed Patches	70.4	60.7	74.6	70.3	77.4	40.9	50.0	42.9	72.1

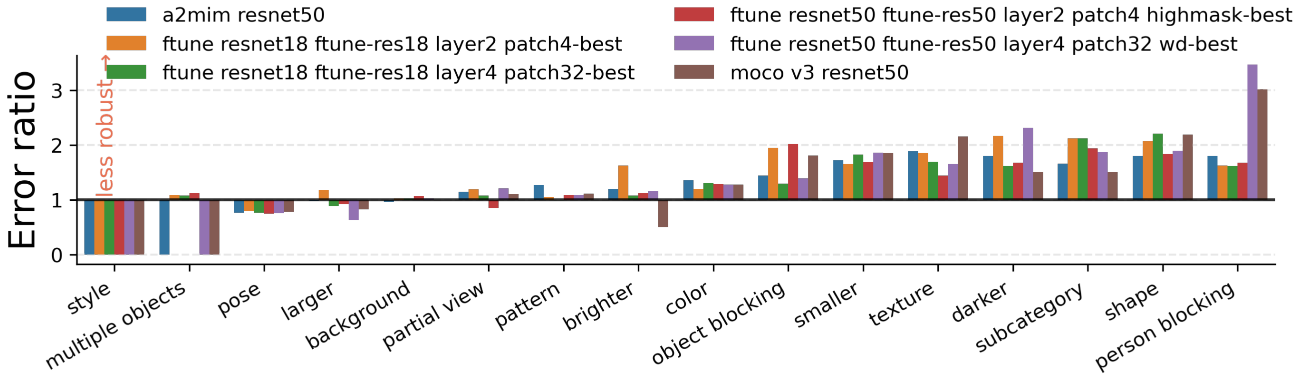

Figure 2. Comparison of model error ratios on ImageNet-X

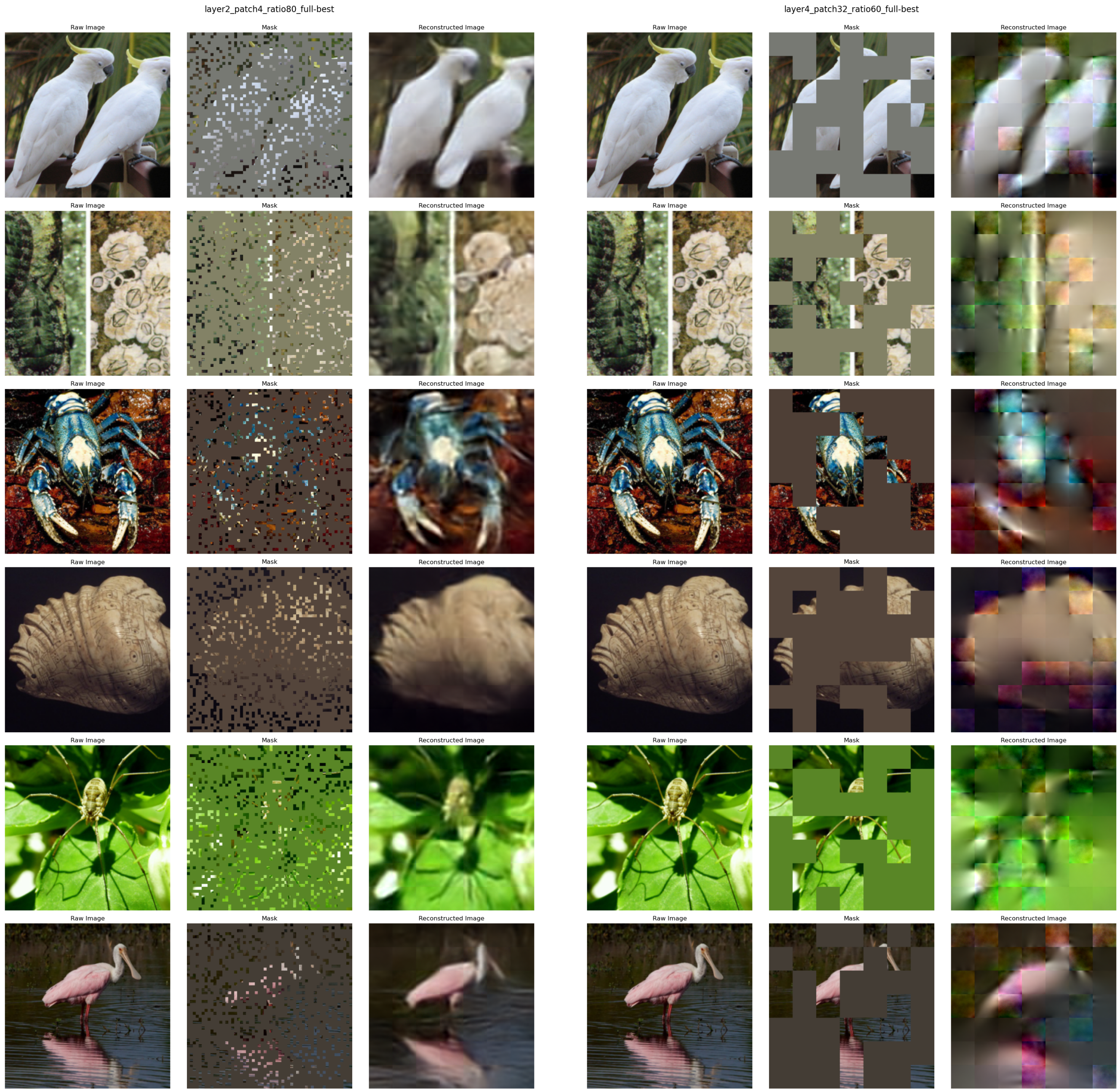

Figure 3. Left: Patch 4 and Mask Ratio 80. Right: Patch 32 and Mask Ratio 60

7. Contributions & Acknowledgements

C.P., J.Y., and S.X. designed and implemented the modified A2MIM system. C.P. trained the neural networks. C.P., J.Y., and S.X. wrote the paper.

References

[1] H. Bao, L. Dong, S. Piao, and F. Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[2] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers, 2021. 1, 2

[3] X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers, 2021. 2

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[5] 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, 2021. 1, 2

[6] 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*, pages 16000–16009, 2022. 2

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners, 2021. 1

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015. 3

[9] B. Y. Idrissi, D. Bouchacourt, R. Balestriero, I. Evtimov, C. Hazirbas, N. Ballas, P. Vincent, M. Drozdzal, D. Lopez-Paz, and M. Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations, 2022. 3

[10] S. Li, D. Wu, F. Wu, Z. Zang, S. Li, et al. Architecture-agnostic masked image modeling–from vit back to cnn. *arXiv preprint arXiv:2205.13943*, 2022. 2, 3

[11] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 1, 2

[12] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[15] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. 2