Is Large-scale Pre-training always Necessary for Vision Transformers?

Alaaeldin El-Nouby[*,1,2] Gautier Izacard[*,1,2] Hugo Touvron[1,3] Ivan Laptev[2]
Hervé Jégou[1] Edouard Grave[1]

[1]Meta AI [2]Inria [3]Sorbonne University

Abstract

Generic large-scale image datasets are powerful means for training visual models. Such datasets, however, often come with limitations. For example, ImageNet has restrictions for commercial usage while automatically crawled large-scale image data may contain unknown biases affecting final models. In this work, we investigate the possibility of achieving a competitive self-supervised pre-training using limited training data available for the target task. We consider datasets such as Stanford Cars, Food101 and COCO, which are order(s) of magnitude smaller than ImageNet. We show that denoising autoencoders, such as BEiT or its variant that we introduce in this paper, are more robust to the type and size of the pre-training data compared to popular self-supervised contrastive learning approaches. We obtain competitive performance compared to ImageNet pre-training for a variety of visual tasks and domains. In particular, for object detection and instance segmentation tasks in COCO, our method outperforms ImageNet pre-trained models, while solely using COCO images for training.

1. Introduction

Modern computer vision neural networks are heavily parametrized: they routinely have tens or hundreds of millions of parameters [1, 2, 3, 4]. This has been the key to their success for leveraging large-scale image collections such as ImageNet. However these high capacity models tend to overfit on small, or even medium sized datasets consisting of hundreds of thousands of images.

The dominant learning paradigm [5, 6] for data-starving problems nowadays is typically: (1) pre-train a model on a large dataset like Imagenet [7], and in turn (2) finetune the weights of the models on the target task for which we have a limited amount of data. The second training stage typically adopts a shorter optimization procedure than the one employed when training from scratch (*i.e.*, from randomly generated weights).

This simple approach has led to impressive results, which are state-of-the-art in many tasks such as detection [8, 9], segmentation [10] and action recognition [11]. Despite this success, we point out that there are some limitations to the reliance on pre-training with curated large-scale datasets. First, most datasets are restricted in terms of their usability in commercial systems as is the case for ImageNet [7][1]. Second, controlling the bias and privacy concerns when dealing with large-scale and web-crawled datasets is challenging. Therefore, it can be advantageous if a method can retain the strong performance of pre-training with large-scale datasets while providing an improved control over copyrights, biases and privacy risks by leveraging smaller sized datasets.

In supervised pre-training, the network learns to focus on the mapping between images and the labels of the pre-training stage, but can discard information that is relevant to other downstream tasks. In other terms, pre-training on large-scale classification datasets does not necessarily align with the goal of learning general-purpose features, as it uses only a subset of the available information controlled by the given dataset categorization bias [12]. These limitations have motivated the development of self-supervised pre-training methods that learn from data without relying on annotations. Most notably, the contrastive and joint embedding approaches [13, 14, 15, 16, 17] can serve as effective pre-training strategies. While obtaining a strong performance on numerous tasks, such methods have a strong bias towards ImageNet data since the transformations have been hand-designed to perform well on the ImageNet benchmark. Some of the most effective transformations, like cropping, rely on the images being object centric [18]. When applied on uncurated data, these methods degrade significantly and require larger datasets to preserve performance [19].

[*]equal contribution

[1]Terms of access explicitly mention *"Researcher shall use the Database only for non-commercial research and educational purposes."* https://image-net.org/download.php

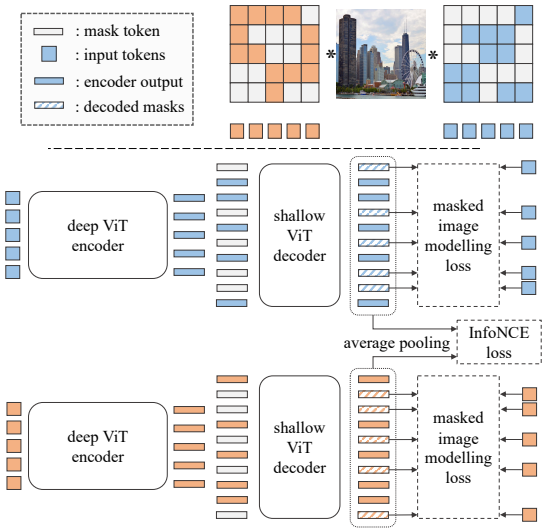

Figure 1. SplitMask process two disjoint subsets of an image independently followed by a a shallow decoder which solves a MIM task for the missing patches in addition to a contrastive signal between two different reconstructions of the same image.

This is in contrast with natural language processing, where nowadays, most applications use large models which were pre-trained on uncurated data. In particular, the (masked) language modeling loss has been applied to transformer networks, leading to the BERT model [20], which is now the foundation of most NLP models. Inspired by this success, Bao et al. [21] have shown the potential of the Masked Image Modeling (MIM) task to pre-train a vision transformer (ViT). Such a model can be thought of as a denoising autoencoder [22] where the noise corresponds to the patch masking operation. This technique has been successfully applied to ImageNet, but research questions remain:

(1) How much does this pre-training method rely on the number of pre-training samples. Does it require millions of images to be useful?
(2) Is this approach robust to different distributions of training images? In particular, is it an effective paradigm to learn with non object-centric or uncurated images?

If the answer to both questions is positive, it will enable pre-training using a larger variety of datasets, including the training sets of many tasks that are smaller or belong to a different domain than ImageNet.

2. Related Work

Pre-training with autoencoders has a long history in deep learning, where it was initially used as a greedy layer-wise method to improve optimization [22, 23, 24, 25, 26]. In the context of unsupervised feature learning for image classification, different tasks related to denoising autoencoders have been considered, such as in-painting [27], colorization [28] or de-shuffling of image patches [29]. In

Table 1. Analysis of different self-supervision methods transfer performance to the iNaturalist-2019 dataset when varying the size of the ImageNet subset used in the pre-training stage, in addition to using non object-centric datasets.

Method	IMNet 1% *epochs: 30k*	IMNet 10% *epochs: 3k*	IMNet Full *epochs: 300*	COCO *epochs: 3k*
Supervised	71.6	75.0	75.8	-
DINO [15]	70.1	73.1	**78.4**	71.9
BEiT [21]	74.1	74.5	75.2	74.4
SplitMask	**74.8**	**75.4**	75.4	**76.3**

Table 2. Ablation study on the effect of different tokenization methods.

	DALL-E	Rand. Proj.	Rand. Patches	K-Means
iNat19	75.2	75.2	75.3	75.0

natural language processing, denoising autoencoders have been applied by masking or randomly replacing some tokens of the input, and reconstructing the original sequence, leading to the BERT model [20]. Similar methods have been proposed to pre-train sequence-to-sequence models, by considering additional kind of noises such as word shuffling or deleting [30, 31].

There has been efforts to adopt such successful ideas in NLP to computer vision, but with limited success. Chen et al. [32] proposed iGPT, a transformer-based autoregressive model that operates over image pixels, while Atito et al. [33] trained a ViT model on denoising of images where the noise is applied at pixel level. More recently, Bao et al. [21] introduced the Masked Image Modeling loss in computer vision, where image patches are masked, and the goal is to predict the discretized label of the missing patches corresponding to their visual words as defined by a pre-trained discrete VAE [34].

Pre-training data is an important ingredient of self-supervised learning, and multiple works have studied its impact on the transfer performance of models. While it is possible to learn high quality features from non-curated (eg. YFCC or IG) data using instance discrimination, this usually requires order of magnitude more data than ImageNet [19, 35]. Similarly, one can perform supervised pre-training using weakly supervised data, such as using hashtags as labels, but this strategy also requires large amount of data to work well [2, 36, 37]. On the other hand, it was shown that for many natural language processing tasks, increasing the size of the pre-training dataset did not lead to strong improvement when using denoising autoencoders [30]. Finally, some work studied how much could be learned from a single pre-training image [38] or from synthetic data [39, 40].

Table 3. COCO detection and instance segmentation performance, using a Mask R-CNN pipeline.

Method	Pre-training			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
	Supervised	IMNet	COCO						
Random Init.	✗	✗	✗	38.3	60.1	41.4	35.6	57.1	37.7
Random Init.†	✗	✗	✗	42.8	64.5	45.6	39.1	61.5	41.7
DeiT [41]	✓	✓	✗	44.2	66.6	47.9	40.1	63.2	42.7
BEiT [21]	✗	✓	✗	44.5	66.2	48.8	40.3	63.2	43.1
DINO [15]	✗	✗	✓	43.7	65.5	47.7	39.6	62.3	42.3
BEiT	✗	✗	✓	44.7	66.3	48.8	40.2	63.1	43.2
SplitMask	✗	✗	✓	**45.3**	**66.9**	**49.4**	**40.6**	**63.6**	**43.5**

Table 4. Finetuning performance on ImageNet. Here, epochs refer to the number of pre-training epochs on ImageNet.

Method	Backbone	Epochs	Top-1
MocoV3 [42]	ViT-S	300	81.4
DINO [15]		300	**81.5**
BEiT [21]		300	81.3
SplitMask		300	**81.5**

3. Analysis

3.1. Sample Efficiency

Denoising autoencoders vs Supervised/DINO First, we start by studying the impact of the pre-training dataset size, by varying the number of ImageNet examples we use to train models. We consider subsets of ImageNet containing 10% and 1% of the total number of examples, and use the balanced (in terms of classes) subsets from [43]. To decouple the effect of using smaller datasets and the effect of doing less training updates, we adapt the number of epochs to keep the number of iterations constant. This means that we perform 3k and 30k epochs on ImageNet 10% and 1% respectively. We report results in Table 1. Observe how pre-training with an autoencoder loss such as masked image modeling is robust to the reduction in dataset size. In contrast, like for supervised pre-training, the performance of models pre-trained with DINO self-supervision degrades when training with smaller datasets.

3.2. Learning using non object-centric images

We now study the impact of changing the nature of the pre-training data. In particular we use images that are not object-centric, like in Imagenet. To this end, instead of pre-training using ImagetNet, we pre-train with images from the COCO dataset only. As COCO contains roughly 118k images, this dataset is approximately equivalent in terms of size to the ImageNet 10% subset. Again, to disentangle the effect of training with a different number of iterations, we adapt the number of epochs: we use 3k epochs on COCO.

We report the results of this experiments in Table 1. When pre-trained on COCO, DINO drops significantly compared to full ImageNet pre-training (-8.3). Interestingly, the drop is higher than using 10% ImageNet even though the numbers of samples is roughly the same. We hypothesis this is because COCO images are not biased to be object-centric, while this joint embedding method was designed and developed using ImageNet as benchmark. In contrast, BEiT's performance only decreases slightly while SplitMask attains +0.7 improvement over full ImageNet pre-training. This is an interesting property which makes such models prime candidates for learning effectively from uncurated images in the wild.

3.3. Tokenizers

The BEiT method, as proposed by Bao et al. [21], relies on the discrete VAE tokenizer from DALL-E, which has been pretrained on a large weakly supervised dataset. Since we want to study whether it is possible to pre-train models solely on small datasets, or non object-centric ones, we replace the DALL-E tokenizer by a simple alternative. To this end, we consider different simple alternatives to discretize images at the patch level without any pre-training as shown in Table 2. Each of these techniques is applied on each patch independently, making them relatively lightweight and more efficient than the original tokenizer considered in BEiT. We observe that replacing the DALL-E tokenizer by simpler choices does not lead to any significant degradation in accuracy. We use *random projection* as our default tokenization method.

4. Methodology

We introduce SplitMask, a variant of denoising autoencoders based on vision transformers. An overview of our method is illustrated in Figure 1.

Our approach is based on three steps, which we refer to as *split*, *inpaint* and *match*. As in standard vision transformers, an image is first broken down into patches of 16×16 pixels. Then, we *split* the patches into two disjoint subsets $\mathcal{A}$ and $\mathcal{B}$, which are processed independently by our deep ViT encoder. Next, using the patch representations of the subset $\mathcal{A}$ and a shallow decoder (e.g. 2 layers), we *inpaint*[2] the patches of the subset $\mathcal{B}$, by solving a MIM task, and vice versa. Finally, we obtain a global image descriptor by average pooling of the patch representations from the decoder output corresponding to each branch.

The feature aggregation is over both observed and hallucinated patches. We try to *match* the global descriptors of the image obtained from subset $\mathcal{A}$ to that obtained from subset $\mathcal{B}$. In other words, we use the masking operation of the mask image modeling loss as a data augmentation for a contrastive learning loss similar to NPID or SimCLR. Note, SplitMask does not add any significant computational cost over MIM methods like BEiT to produce this global con-

[2]Inpainting in this context is implemented by solving a Masked Image Modeling task rather than the typical inpainting by reconstruction of pixels.

Table 5. Comparison between finetuning performance on the target datasets when different pre-training datasets are used.

Method	Backbone	Supervised pre-training	Data Used		iNat-18	iNat-19	Food 101	Cars	Clipart	Painting	Sketch
			IMNet	Target	437k	265k	75k	8k	34k	52k	49k
Liu et al. [44]∓	CVT-13	✗	✗	✓	-	-	-	-	60.6	55.2	57.6
	ResNet-50	✗	✗	✓	-	-	-	-	63.9	53.5	59.6
Random Init.	ViT-S	✗	✗	✓	59.6	67.5	84.7	35.3	41.0	38.4	37.2
DeiT [41]		✓	✓	✓	69.9	75.8	**91.5**	92.2	**79.6**	**74.2**	**72.5**
BEiT [21]		✗	✓	✓	68.1	75.2	90.5	92.4	75.3	68.7	68.5
BEiT		✗	✗	✓	68.8	76.1	90.7	92.7	-	69.0	-
SplitMask		✗	✗	✓	**70.1**	**76.3**	**91.5**	**92.8**	78.3	69.2	70.7

trastive training signal.

5. Experiments

5.1. Object detection and Instance Segmentation

First, we evaluate our approach on the COCO object detection and instance segmentation dataset using the Mask R-CNN pipeline [8] and report our results in Table 3. We compare models pre-trained on the COCO dataset alone with their equivalent counterparts that were pre-trained on ImageNet, either in a supervised or self-supervised fashion. First, we observe that BEiT models which were pre-trained on the COCO dataset alone obtain better downstream task performance than the same models pre-trained on ImageNet. For example, when using a ViT-base backbone, pre-training on COCO instead of ImageNet leads to a boost of +0.4 in box AP.

Finally, we observe that SplitMask leads to a consistent improvement compared to the BEiT baseline, such as +0.6 box AP when using a ViT-small and +0.3 mask AP for ViT-base backbones. All put together, in a comparable setting, we obtain a +1.1 box AP increase while not using ImageNet.

5.2. Image Classification

We perform empirical evaluation on a number classification datasets and report our results in Table 5.

BEiT pre-training: ImageNet vs Target First, we compare ImageNet pre-training to the target data pre-training with BEiT and observe that for many cases, pre-training on the target data alone leads to better results. This is true for the ViT-small backbone across all the datasets including Stanford cars (+1.1% acc), which consists of only 8k images. When using a ViT-base backbone, pre-training on the target task data outperforms BEiT self-supervised ImageNet pre-training for datasets as small as Food101 (+0.7 acc), which is more than 10x smaller than ImageNet. Second, we observe that SplitMask leads to further improvement in performances for multiple datasets: for example, on the iNaturalist 2018 dataset, we see +3.0 in accuracy with a ViT-base model.

Supervised ImageNet pre-training As it was already observed in previous work [15, 16, 42], we also see in many cases that self-supervised training outperforms supervised pre-training on ImageNet. For example, on the iNaturalist datasets, training with the target task data alone (including a pre-training step) gives better results than pre-training on ImageNet with labels: with a ViT-base model and the SplitMask method, we see an improvement of +2.7% in top-1 accuracy. As for the *clipart*, *painting* and *sketch* datasets, we see that SplitMask provides a competitive performance, outperforming an ImageNet pre-trained BEiT across all datasets for ViT-S. However, for the aforementioned datasets, supervised pre-training achieves the best performance for both ViT-S and ViT-B.

5.3. Pre-training using ImageNet

In Table 4 we show the performance of our SplitMask method using the ViT-S and ViT-B backbones and 300 epochs pre-training compared to other recent transformer-based self-supervised learning methods. It can be observed that SplitMask provides a strong performance, outperforming both BEiT and MocoV3 for all backbones. Additionally, SplitMask achieves a performance on par with DINO while being significantly cheaper and simpler to train. Note that while SplitMask and BEiT attain a strong finetuning performance, denoising autoencoding methods typically fall behind in terms of linear probing compared to instance discrimination methods like DINO.

6. Conclusion

In this paper, we have raised the question of how to pre-train models with self-supervised learning, wondering in particular on whether large scales datasets such as Imagenet are necessary for pre-training. Our study on ImageNet shows that taking a smaller pre-training dataset does not lead to big performance drop for denoising autoencoders, as opposed to instance discrimination self-supervised techniques or supervised pre-training. Similarly, training on non object-centric images does not impact the downstream task performance significantly.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016. 1

[2] 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 *International Conference on Learning Representations*, 2021. 1, 2, I

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021. 1, I

[4] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár, "Designing network design spaces," in *Computer Vision and Pattern Recognition*, 2020. 1

[5] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Computer Vision and Pattern Recognition*, 2014. 1

[6] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson, "How transferable are features in deep neural networks?" *arXiv preprint arXiv:1411.1792*, 2014. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*, 2009. 1

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *International Conference on Computer Vision*, 2017. 1, 4, I

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020. 1

[10] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014. 1

[11] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308. 1

[12] Eleanor H. Rosch, "Natural categories," *Cognitive Psychology*, 1973. 1

[13] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Computer Vision and Pattern Recognition*, 2020. 1

[14] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *arXiv preprint arXiv:2006.09882*, 2020. 1

[15] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," *arXiv preprint arXiv:2104.14294*, 2021. 1, 2, 3, 4

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020. 1, 4

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020. 1

[18] Senthil Purushwalkam and Abhinav Gupta, "Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases," *arXiv preprint arXiv:2007.13916*, 2020. 1

[19] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin *et al.*, "Self-supervised pretraining of visual features in the wild," *arXiv preprint arXiv:2103.01988*, 2021. 1, 2

[20] 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. 2

[21] Hangbo Bao, Li Dong, and Furu Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021. 2, 3, 4, I

[22] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103. 2

[23] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Re-

ducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006. 2

[24] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, "Greedy layer-wise training of deep networks," in *Advances in neural information processing systems*, 2007, pp. 153–160. 2

[25] Marc Ranzato, Christopher Poultney, Sumit Chopra, Yann LeCun *et al.*, "Efficient learning of sparse representations with an energy-based model," *Advances in neural information processing systems*, vol. 19, p. 1137, 2007. 2

[26] 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*, vol. 11, no. 12, 2010. 2

[27] 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*, 2016, pp. 2536–2544. 2

[28] Richard Zhang, Phillip Isola, and Alexei A Efros, "Colorful image colorization," in *European conference on computer vision*. Springer, 2016, pp. 649–666. 2

[29] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84. 2

[30] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019. 2

[31] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019. 2

[32] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever, "Generative pretraining from pixels," in *International Conference on Machine Learning*, 2020. 2

[33] Sara Atito, Muhammad Awais, and Josef Kittler, "Sit: Self-supervised vision transformer," *arXiv preprint arXiv:2104.03602*, 2021. 2

[34] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021. 2

[35] Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin, "Unsupervised pre-training of image features on non-curated data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2959–2968. 2

[36] Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache, "Learning visual features from large weakly supervised data," in *European Conference on Computer Vision*. Springer, 2016, pp. 67–84. 2

[37] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 181–196. 2

[38] Yuki M Asano, Christian Rupprecht, and Andrea Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," *arXiv preprint arXiv:1904.13132*, 2019. 2

[39] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh, "Pre-training without natural images," in *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[40] Kundan Krishna, Jeffrey Bigham, and Zachary C Lipton, "Does pretraining for summarization require knowledge transfer?" *arXiv preprint arXiv:2109.04953*, 2021. 2

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, "Training data-efficient image transformers and distillation through attention," *arXiv preprint arXiv:2012.12877*, 2020. 3, 4

[42] Xinlei Chen, Saining Xie, and Kaiming He, "An empirical study of training self-supervised vision transformers," *arXiv preprint arXiv:2104.02057*, 2021. 3, 4, I

[43] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat, "Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples," *arXiv preprint arXiv:2104.13963*, 2021. 3

[44] Yahui Liu, Enver Sangineto, Wei Bi, Nicu Sebe, Bruno Lepri, and Marco Nadai, "Efficient training of

visual transformers with small datasets," in *Advances in Neural Information Processing Systems*, 2021. 4

[45] 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. I

[46] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Computer Vision and Pattern Recognition*, 2017. I

[47] Alaaeldin El-Nouby, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *arXiv preprint arXiv:2106.09681*, 2021. I

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017. I