Fake & Square: Training Self-Supervised Vision Transformers with Synthetic Data and Synthetic Hard Negatives

Nikolaos Giakoumoglou Andreas Floros Kleanthis Marios Papadopoulos Tania Stathaki
Imperial College London
{nikos, andreas.floros18, kleanthis-marios.papadopoulos18, t.stathaki}@imperial.ac.uk

Abstract

This paper does not introduce a new method per se. Instead, we build on existing self-supervised learning approaches for vision, drawing inspiration from the adage "fake it till you make it". While contrastive self-supervised learning has achieved remarkable success, it typically relies on vast amounts of real-world data and carefully curated hard negatives. To explore alternatives to these requirements, we investigate two forms of "faking it" in vision transformers. First, we study the potential of generative models for unsupervised representation learning, leveraging synthetic data to augment sample diversity. Second, we examine the feasibility of generating synthetic hard negatives in the representation space, creating diverse and challenging contrasts. Our framework—dubbed Syn^{2}Co—combines both approaches and evaluates whether synthetically enhanced training can lead to more robust and transferable visual representations on DeiT-S and Swin-T architectures. Our findings highlight the promise and limitations of synthetic data in self-supervised learning, offering insights for future work in this direction.

1. Introduction

The field of computer vision has witnessed remarkable progress in recent years. The progress of AI systems has been fueled by large-scale datasets that are carefully annotated by humans. However, the reliance on manually labeled data has become a bottleneck for further advancements, leading to increased interest in self-supervised learning methods that leverage vast amounts of unlabeled data [18]. Deep learning models are particularly data-hungry, requiring vast amounts of labeled data to perform well, yet the rapid expansion of available data has outpaced our capacity to label it comprehensively. Contrastive methods [6, 14] have demonstrated that self-supervised learning can rival and even surpass supervised pre-training, particularly in tasks requiring robust feature representations.

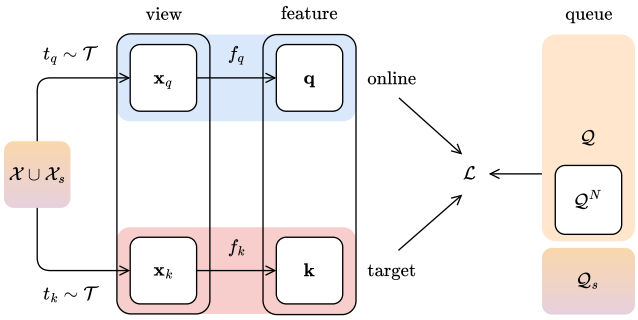


Figure 1. Overview of our pipeline for contrastive learning. It includes synthetic data, $\mathcal{X}_s$, produced by a diffusion model and synthetic negatives, $\mathcal{Q}_s$, which are generated on the fly.

Despite their success, contrastive learning methods face two significant challenges. First, they often struggle with the limited diversity of hard negative examples, which are crucial for learning discriminative features [16, 31]. Second, they typically require vast amounts of real-world data to achieve competitive performance, creating a data bottleneck similar to supervised learning, limiting scalability, generalization, and applicability to low-resource domains [11].

We are inspired by the adage: *"fake it till you make it"* – presenting something "*synthetic*" as if it were "*real*" until it actually produces results. We apply this concept to unsupervised contrastive learning in two ways: by introducing synthetic data clones in the image space, effectively expanding the training dataset, and by generating synthetic hard negatives in the representation space, creating challenging contrasts. We focus on vision transformers, which have achieved remarkable success, surpassing ResNets [8].

The main **contributions** of our work are as follows:

- We create a synthetic clone of ImageNet-100 and demonstrate how synthetic data affects contrastive self-supervised learning.
- We adapt and extend synthetic negative generation to transformers, showing that DeiT and Swin can leverage synthetic contrasts to learn more discriminative features.

Figure 2. DeiT Grad-CAMs. Left: input image. Middle: pretraining with real data. Right: pretraining with synthetic data.

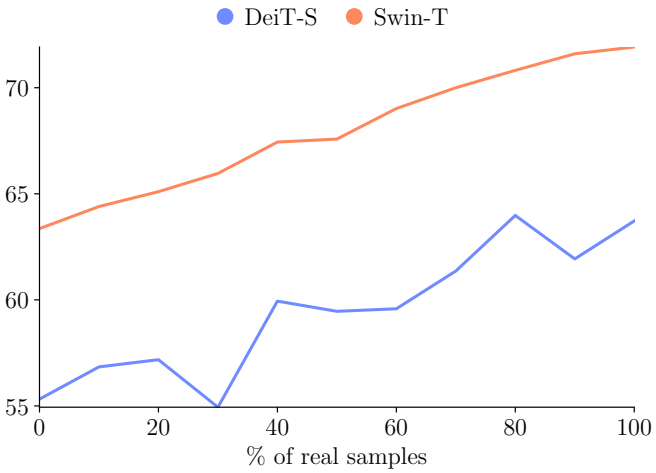


Figure 3. Top-1 linear probing accuracy on ImageNet-100. A percentage of the pretraining data is real and the rest is synthetic.

2. Related Work

Self-supervised visual representation learning. Self-supervised learning has improved visual representation learning through various paradigms [11]. Recent methods include contrastive learning [6, 14], clustering [3, 4], and distillation approaches [5, 13]. While SimCLR [6] established the importance of augmentation strategies and large batch sizes, MoCo [14] introduced memory banks for scalable negative sampling. The quality of negative samples in contrastive learning has been a focus of extensive research [1, 12, 16]. These studies aim to select informative negative samples and address false negatives in instance discrimination tasks.

Transformers for vision. The success of Transformers [27] in NLP has inspired continuous efforts to generalize this architecture to computer vision tasks [10, 19, 25]. ViT [10] demonstrated that a pure transformer applied to sequences of image patches can excel in image classification tasks, particularly when pre-trained on large datasets. DeiT [25] further showed that ViTs can be effectively trained on ImageNet without requiring external data. Swin Transformer [19] introduced a hierarchical design with shifted windows, making it more suitable for dense prediction tasks. Self-supervised learning for vision transformers has rapidly evolved, including contrastive adaptations [7, 29], masked image modeling methods [15], and distillation frameworks [5, 32].

Synthetic data for vision. Synthetic data generation has become increasingly important in computer vision [20]. This field has advanced significantly through GANs that produce highly realistic images [17]. Recent research has explored generating synthetic data specifically for ImageNet classes [9] using class-conditional GANs [2]. Beyond generation, self-supervised learning methods have effectively leveraged synthetic data, with SynCLR [24] demonstrating that learning from synthetic images can rival results obtained from real data. Similarly, Sariyildiz et al. [21] showed that models trained on diffusion-generated ImageNet clones can perform comparably to those trained on real data, exhibiting strong generalization capabilities.

3. Methodology

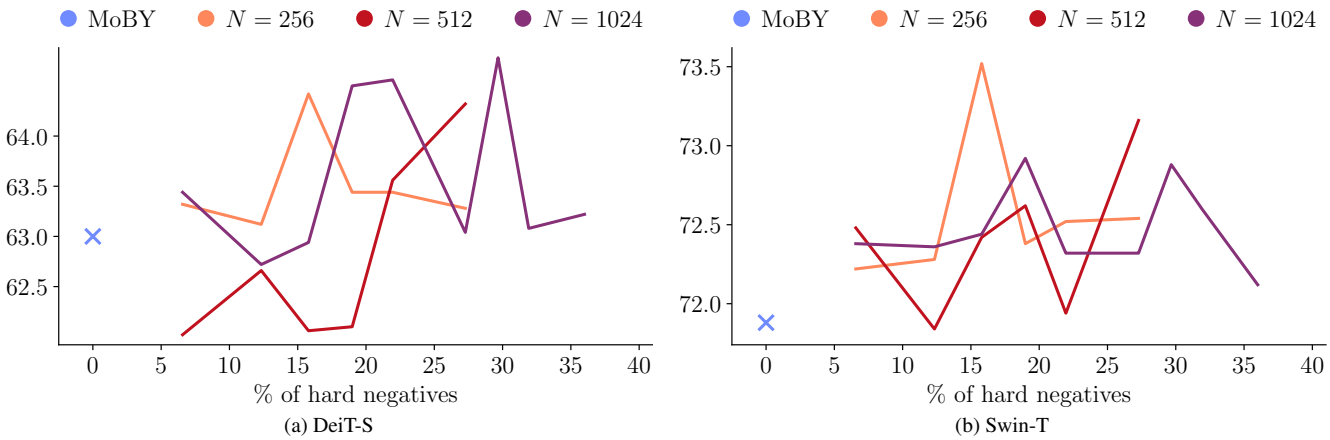
In this section, we present our approach to incorporating synthetic elements into self-supervised learning. We formulate the task (Section 3.1), describe the contrastive learning framework (Section 3.2), and detail our methods for generating synthetic data (Section 3.3) and synthetic hard negatives (Section 3.4). We illustrate our complete method, which we refer to as Syn^{2}Co, in Figure 1.

3.1. Task Formulation

Our goal is to learn meaningful representations without relying on labeled data or purely real image datasets. Formally, given a dataset of images, $\mathcal{X}$, we aim to learn a representation function, $f_q : \mathcal{X} \to \mathcal{Z}$, that produces a feature space where semantically similar images are close together and dissimilar images are far apart. Unlike standard self-supervised learning, here we explore the possibility of enhancing the training process by augmenting the data with synthetic images, $\mathcal{X}_s$, synthesized using a generative diffusion model $\mathcal{G}$. Additionally, we use synthetic hard negatives, $\mathcal{Q}_s$, constructed in the representation space.

3.2. Contrastive Learning

We employ a contrastive approach to learn representations by comparing similar (*positives*) and dissimilar (*negatives*) samples. Given an image, $\mathbf{x} \sim \mathcal{X}$, and a family of image



(a) DeiT-S (b) Swin-T

Figure 4. Top-1 linear probing accuracy on ImageNet-100. We pretrain for 100 epochs on real data (*i.e.*, ImageNet-100 training set) using different hardness selection values and synthetic negative percentages.

augmentations, $\mathcal{T}$, we create two views of the same image: $\mathbf{x}_q = t_q(\mathbf{x})$ and $\mathbf{x}_k = t_k(\mathbf{x})$, where $t_q, t_k \sim \mathcal{T}$. These views are then mapped using an *online* encoder (our desired representation function), f_q, and a *target* encoder, f_k, producing feature vectors $\mathbf{q} = f_q(\mathbf{x}_q)$ and $\mathbf{k} = f_k(\mathbf{x}_k)$. f_q is optimized via backpropagation, by minimizing the following InfoNCE [26] contrastive objective:

$$\mathcal{L} = -\log \frac{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau)}{\exp(\mathbf{q}^\top \cdot \mathbf{k}/\tau) + \sum_{\mathbf{n}\in\mathcal{Q}} \exp(\mathbf{q}^\top \cdot \mathbf{n}/\tau)}, \quad (1)$$

where $\mathcal{Q} = \{\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_K\}$ is a set of K negative samples and τ is the temperature that controls sharpness. Negative samples are either drawn from the current batch [6] or from a memory queue [14, 28]. f_k mostly mirrors f_q and is updated via a momentum mechanism at each step:

$$\theta_k \leftarrow m \cdot \theta_k + (1 - m) \cdot \theta_q. \quad (2)$$

3.3. Synthetic Data

We use an image generator, $\mathcal{G}$, to create a synthetic dataset, $\mathcal{X}_s$, that complements the real dataset $\mathcal{X}$. Specifically, we leverage the diffusion model from [22] to generate clones of ImageNet-100. This approach results in a synthetic dataset of approximately 130,000 images that can be used either as a replacement for or in conjunction with the real dataset. During training, images may be sampled from $\mathcal{X} \cup \mathcal{X}_s$. This mixed sampling strategy enables us to investigate various training scenarios, including training exclusively on real data ($\mathcal{X}$ only), training exclusively on synthetic data ($\mathcal{X}_s$ only), and training on a mixture of real and synthetic data with different mixing ratios (see Figure 2 and Figure 3).

3.4. Synthetic Hard Negatives

We generate synthetic hard negatives directly in the representation space [12, 16, 30] to provide challenging examples that help the model learn more discriminative features. This addresses a key limitation in contrastive learning, *i.e.*, the scarcity of informative negative examples that effectively capture semantic boundaries. Given a query representation, $\mathbf{q} \in \mathcal{Z}$, and a negative queue, $\mathcal{Q}$, we first select the top N hardest negatives, $\mathcal{Q}^N \subseteq \mathcal{Q}$, based on their similarity to $\mathbf{q}$:

$$\mathcal{Q}^N = \text{TopK}(\{\text{sim}(\mathbf{q}, \mathbf{n}) \mid \mathbf{n} \in \mathcal{Q}\}, N), \quad (3)$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) := \mathbf{a}^\top \cdot \mathbf{b}/(\|\mathbf{a}\|_2 \cdot \|\mathbf{b}\|_2)$ is the cosine similarity and $\|\cdot\|_2$ denotes the ℓ_2 norm. These hardest negatives serve as the basis for constructing synthetic contrasts that the model must learn to differentiate. We formulate synthetic hard negatives through a synthesis function, $\mathcal{F} : \mathcal{Z} \times \mathcal{Z} \to \mathcal{Z}$, that operates on the query and selected hard negatives:

$$\mathbf{s} = \frac{\mathcal{F}(\mathbf{q}, \mathbf{n})}{\|\mathcal{F}(\mathbf{q}, \mathbf{n})\|_2}, \quad \mathbf{n} \in \mathcal{Q}^N, \quad (4)$$

where $\mathbf{s}$ is the normalized synthetic negative. Following [12], we implement six different synthesis strategies for $\mathcal{F}$, including interpolation, extrapolation, mixing, noise jittering, perturbation, and adversarial methods.

The set of L synthetic negatives, $\mathcal{Q}_s = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L\}$, is then combined with the existing queue of real negatives, $\mathcal{Q}$, effectively expanding the diversity of negative examples. This combination of real and synthetic negatives allows the model to learn more robust and generalizable representations by providing more diverse and challenging contrasts during training, without requiring additional real data collection.

Table 1. Top-1 and top-5 linear probing accuracy on ImageNet-100 for different methods. We reproduce all methods.

Method	Synthetic Data	Synthetic Negatives	Epochs	Iterations	DeiT-S		Swin-T	
					Top-1 (%)	Top-5 (%)	Top-1 (%)	Top-5 (%)
Supervised	-	-	300	76,200	80.50	94.20	86.78	97.14
DINO [5]†	-	-	300	304,800	75.42	93.82	-	-
MoCo-v3 [7]‡	✗	✗	400	25,600	79.41	92.20	-	-
MoBY [29]	✗	✗	300	76,200	79.36	95.32	83.90	96.76
Syn^{1}Co*	✗	✓	300	76,200	78.96	95.06	**84.04**	**96.94**
	✓	✗	150	76,200	79.02	94.94	83.45	96.01
	✓	✗	300	152,400	81.86	95.80	83.68	96.84
Syn^{2}Co	✓	✓	150	76,200	78.12	94.64	83.55	96.24
	✓	✓	300	152,400	**82.12**	**95.86**	83.70	96.76

† Trained with batch size 128. ‡ Trained with batch size 2048.

* When including either synthetic data or synthetic negatives (but not both), we refer to this as Syn^{1}Co for fun.

4. Experiments

We present comprehensive comparisons with state-of-the-art methods (Section 4.2) and subsequently conduct extensive experiments to evaluate our approach. We examine how synthetic data, $\mathcal{X}_s$, influences model performance (Section 4.3) and analyze the impact of synthetic hard negatives, $\mathcal{Q}_s$, on representation quality (Section 4.4).

4.1. Setup

We build on the implementations of MoBY [29] and SynCo [12]. Specifically, we use MoBY as a baseline with the DeiT-Small [25] and Swin-Tiny [19] architectures, following the same augmentation strategy as BYOL [13]. For synthetic negative generation, we implement all six strategies from SynCo [12] and explore configurations with $N \in \{256, 512, 1024\}$ hardest negatives and synthetic ratios $L/(K+L)$. We conduct our experiments on ImageNet-100 [9, 23]. For our synthetic data experiments, we generate a matching set of 130,000 images using the diffusion model from [22], maintaining the same class distribution as the original dataset. We evaluate the quality of learned representations on ImageNet-100 classification using linear probing.

4.2. Results

Table 1 presents a comprehensive evaluation of our proposed approach and comparisons with existing methods. Swin benefits primarily from synthetic negatives, achieving optimal performance with synthetic negatives alone. In contrast, DeiT effectively leverages both synthetic components, reaching peak performance with the complete approach during extended training. These architecture-specific responses highlight the importance of considering transformer design when applying synthetic augmentation strategies.

4.3. Effect of Synthetic Data

Figure 3 demonstrates that models learn effective representations from synthetic data alone, though performance improves with increasing real data proportions. Results indicate synthetic images serve as valuable complements rather than complete substitutes for real data. The relatively modest performance gap between fully synthetic and fully real training regimes suggests that modern diffusion models can generate high-quality samples that capture substantial semantic information from the original dataset.

4.4. Effect of Synthetic Negatives

Figure 4 shows architecture-specific responses to synthetic negatives. DeiT exhibits sensitivity to both hardness levels and proportions, while Swin maintains consistent performance gains across configurations.

5. Discussion

Our experiments show that architectural differences significantly impact how models leverage synthetic elements. The inclusion of synthetic data follows a diminishing returns pattern, which aligns with findings that diffusion models produce distribution shifts, limiting their effectiveness as complete substitutes. Synthetic negatives provide a computationally efficient enhancement, though their hyperparameters must be carefully tuned.

Acknowledgments

We acknowledge the computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232), which enabled our experiments.

References

[1] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019. 2

[2] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This dataset does not exist: training models from generated images. In *ICASSP*, 2020. 2

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features, 2019. 2

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924. Curran Associates, Inc., 2020. 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. 2, 4

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 1, 2, 3

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers, 2021. 2, 4

[8] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations, 2022. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2, 4

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[11] Nikolaos Giakoumoglou and Tania Stathaki. A review on discriminative self-supervised learning methods, 2024. 1, 2

[12] Nikolaos Giakoumoglou and Tania Stathaki. Synco: Synthetic hard negatives in contrastive learning for better unsupervised visual representations, 2024. 2, 3, 4

[13] 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, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 2, 4

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. 1, 2, 3

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 2

[16] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020. 1, 2, 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[18] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Liming Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):21–40, 2021. 1

[19] 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, 2021. 2, 4

[20] Sergey I Nikolenko. Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512*, 2021. 2

[21] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[22] Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020. 4

[24] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. *arXiv preprint arXiv:2312.17742*, 2023. 2

[25] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention, 2021. 2, 4

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 3

[27] 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*, pages 5998–6008, 2017. 2

[28] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 3

[29] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-supervised learning with swin transformers, 2021. 2, 4

[30] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning, 2022. 3

[31] Wenzheng Zhang and Karl Stratos. Understanding hard negatives in noise contrastive estimation, 2021. 1

[32] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer, 2022. 2