Rejuvenating image-GPT as Strong Visual Representation Learners

Sucheng Ren [*1] **Zeyu Wang** [*2] **Hongru Zhu** [1] **Junfei Xiao** [1] **Alan Yuille** [1] **Cihang Xie** [2]

Abstract

This paper enhances image-GPT (iGPT), one of the pioneering works that introduce autoregressive pretraining to predict the next pixels for visual representation learning. Two simple yet essential changes are made. First, we shift the prediction target from raw pixels to semantic tokens, enabling a higher-level understanding of visual content. Second, we supplement the autoregressive modeling by instructing the model to predict not only the next tokens but also the visible tokens. This pipeline is particularly effective when semantic tokens are encoded by discriminatively trained models, such as CLIP. We introduce this novel approach as D-iGPT. Extensive experiments showcase that D-iGPT excels as a strong learner of visual representations: A notable achievement is its compelling performance on the ImageNet-1K dataset — by training on publicly available datasets, D-iGPT unprecedentedly achieves **90.0%** top-1 accuracy with a vanilla ViT-H. Additionally, D-iGPT shows strong generalization on the downstream task. Code is available at https://github.com/OliverRensu/D-iGPT.

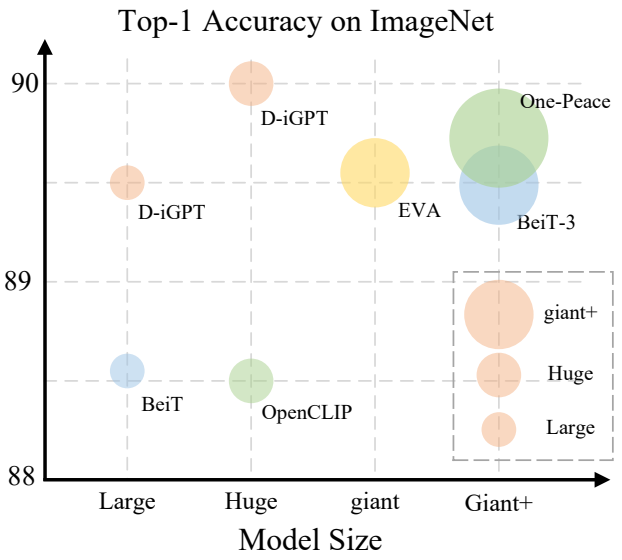


Figure 1. ImageNet performance of models trained on publicly available datasets. We note that D-iGPT with ViT-H achieves the best performance, *i.e.*, 90.0% top-1 accuracy.

1. Introduction

The advent of Large Language Models (LLMs) (OpenAI, 2023; Thoppilan et al., 2022; Touvron et al., 2023), such as GPT series (Radford & Narasimhan, 2018; Brown et al., 2020; OpenAI, 2023), has catalyzed a transformative era in natural language processing (NLP), establishing new precedents for performance across a range of linguistic tasks. One of the key driving forces behind this tremendous success is autoregressive pretraining, which trains models to predict the most probable next tokens in a sequence. This strategy enables models to internalize a complex interplay of syntax and semantics, which in turn translates to their extraordinary prowess to process language with human-like capabilities.

Beyond NLP, autoregressive pretraining has also been a significant contributor in the field of computer vision. The pioneering model in this context is PixelCNN (Van Den Oord et al., 2016), a deep autoregressive model designed to model the discrete probability of the raw pixel values and encode the complete set of dependencies in the image. Building upon this foundation, image GPT (iGPT) (Chen et al., 2020a) represents a significant advancement, utilizing the flexible Transformer architecture (Vaswani et al., 2017) at a notably larger computational scale. iGPT's achievements are remarkable: it not only learned state-of-the-art visual representation for lower-resolution datasets such as CIFAR-10 but also demonstrated competitive performance on more complex datasets like ImageNet (Russakovsky et al., 2015).

Intriguingly, despite the initial successes of autoregressive pretraining in computer vision, recent trends have witnessed a rapid paradigm shift towards BERT-style pretraining (Devlin et al., 2019). This transition is significant, particularly when considering iGPT's initial findings of comparable performance between autoregressive and BERT-style pretraining in various tasks. Subsequent research, however,

[*]Equal contribution [1]Johns Hopkins University [2]UC Santa Cruz. Correspondence to: Cihang Xie <cixie@ucsc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

has increasingly favored BERT-style pretraining (Bao et al., 2022; He et al., 2022) for its superior efficacy in visual representation learning. For example, MAE (He et al., 2022) demonstrates that simply predicting the values of randomly masked pixels can effectively serve as a scalable solution for visual representation learning.

In this paper, we revisit iGPT, challenging that *autoregressive pretraining is actually capable of building strong vision learners, especially at scale*. Our methodology incorporates two critical modifications. First, acknowledging that images are inherently noisy and redundant, we follow BEiT (Bao et al., 2022) to "tokenize" images into semantic tokens. This adjustment reorients the autoregressive prediction focus from pixels to semantic tokens, thereby enabling a more nuanced understanding of the interplay among different image regions. Second, we complement the generative decoder, which is responsible for autoregressively predicting the next semantic token, with a discriminative decoder. This additional component is tasked with predicting the semantic tokens of the visible pixels. Moreover, an intriguing observation is that this pretraining pipeline works best when the semantic visual tokens are derived from models trained discriminatively, such as CLIP (Radford et al., 2021). We term this enhanced approach as D-iGPT.

Extensive experiments across various datasets and tasks confirm the effectiveness of our proposed D-iGPT. With ImageNet-1K as the sole pertaining dataset, our base-size model achieves an 86.2% top-1 classification accuracy, surpassing previous state-of-the-art by 0.6%. By further scaling to the larger ImageNet-21K dataset, our huge-size model unprecedentedly achieves a **90.0%** top-1 classification accuracy, outperforming all existing solutions developed using public datasets. We hope this work can catalyze the community to reevaluate the potential of autoregressive pretraining for visual representation learning.

2. Related Work

2.1. Self-supervised Learning

According to learning targets, self-supervised learning can be labeled as discriminative-based or generative-based.

Discriminative Self-supervised Learning. This paradigm focuses on learning transferable representation by defining a pre-task that scores the discriminative power of learned representations. A notable strategy within this category is contrastive learning, which utilizes a contrastive loss to learn representation similarity or dissimilarity between the same images with different augmentation or entirely different images. For instance, Wu *et al*. (Wu et al., 2018) introduces instance discrimination, constructing positive and negative query-key pairs from the same or different images. SimCLR (Chen et al., 2020b) further improves the performance with a projection head, strong data augmentations, and large-batch-size training. MoCo (He et al., 2020; Chen et al., 2020c) incorporates a memory bank and a momentum encoder without the need for large batch sizes. CLIP (Radford et al., 2021) extends this concept by incorporating language supervision through image-text pairings.

Generatieve Self-supervised Learning. In contrast to the discriminative approaches, generative self-supervised learning emphasizes training models to reconstruct the original inputs from corrupted versions.

Masked image modeling, inspired by BERT (Devlin et al., 2019) in NLP, is the dominant strategy in this line of research. For example, the pioneering work BEiT (Bao et al., 2022) pretrains models to recover the corresponding semantic tokens based on the corrupted image patches. Other significant methods include MAE (He et al., 2022), SimMIM (Xie et al., 2022), MaskFeat (Wei et al., 2021), PeCo (Dong et al., 2021), MILAN (Hou et al., 2022), DeepMIM (Ren et al., 2023a).

This study pivots towards a distinct facet of generative self-supervised learning, namely, autoregressive pretraining. In NLP, autoregressive pretraining is also highly regarded alongside BERT-style methods, especially effective in the era of LLMs (OpenAI, 2023; Touvron et al., 2023). However, its progress in computer vision has not yet paralleled the heightened interest sparked by the initial success of iGPT (Chen et al., 2020a). This paper aims to bridge this gap. We demonstrate that, with simple yet essential modification, autoregressive pretraining exhibits extraordinary capabilities in building strong vision models.

2.2. ImageNet-1K Winning Solutions

The advancements in ImageNet-1K performance have seen a significant boost, primarily driven by scaling datasets and model sizes. Liu *et al*. (Liu et al., 2022b) exemplify this trend with the successful training of SwinV2-G, a model equipped with ∼3 billion parameters, using techniques like residual-post-norm and scaled cosine attention. Similarly, Dehghani *et al*. (Dehghani et al., 2023) have shown the impressive capabilities of ViT-22B, highlighting the feasibility of "LLM-like" scaling in computer vision. Zhang *et al*. (Zhai et al., 2022) investigate scaling both model and data, providing valuable insights into the interplay between scaling factors and performance. Another noteworthy development is by Chen *et al*. (Chen et al., 2023) which discovers deep neural network training algorithms through program search, leading to the creation of the effective and memory-efficient optimizer Lion. However, a common limitation across these methods is their heavy reliance on private, in-house data, such as JFT-3B (Zhai et al., 2022), which raises significant reproducibility concerns.

In contrast to the approaches above, there is a notable trend of employing public datasets to train more powerful vision models. For instance, Wang *et al*. (Wang et al., 2022) scale BEiT-3 to ~1.9 billion parameters using a combination of images, texts, and image-text pairs, all sourced from public datasets. Likewise, Fang *et al*.(Fang et al., 2022) successfully scaled up EVA, a vanilla ViT with ~1 billion parameters, using a total of ~29.6 million public images. One-Peace (Wang et al., 2023) presents a 4-billion-parameter model capable of unifying vision, audio, and language representations. Our D-iGPT model stands out in this landscape by achieving superior performance than EVA and One-Peace, and meanwhile using smaller model and data sizes.

3. Method

We hereby first revisit iGPT in Section 3.1. Next, we present our enhanced version, D-iGPT, in Section 3.2, which shifts the prediction target from raw pixels to semantic tokens and additionally supplies supervision on visible tokens. Lastly, the specifics of our model's architecture, along with implementation details.

3.1. Revisiting iGPT

GPT. In NLP, the generative pretraining involves modeling the probability of the next word in a corpus $\mathcal{U} = \{u_1, ..., u_n\}$ autoregressively. This can be written as:

$$p(u) = \prod_{i=1}^{n} p(u_i | u_1, ..., u_{i-1}, \Theta) \tag{1}$$

Here, GPT computes the likelihood of each word u_i based on the context of all preceding words from u_1 to u_{i-1}, aiming to minimize the negative log-likelihood of the target words:

$$\mathcal{L} = -log\ p(u) \tag{2}$$

Image GPT. In the context of images, where the input is an image $X \in \mathcal{R}^{H \times W \times C}$, the challenge lies in converting this 2D structure into a sequential format akin to a language sequence. iGPT (Chen et al., 2020a) addresses this by naïvely vectorizing the image X into a series of individual pixels $\{x_1, ..., x_n\}$, treating each pixel as analogous to a word. It then models the probability of each subsequent pixel based on the preceding ones in the sequence:

$$p(x) = \prod_{i=1}^{n} p(x_i | x_1, ..., x_{i-1}, \Theta) \tag{3}$$

In this formulation, iGPT aims to predict each pixel x_i utilizing the information from preceding pixels $\{x_1, ..., x_{i-1}\}$, minimizing the negative log-likelihood:

$$\mathcal{L} = -log\ p(x) \tag{4}$$

Nevertheless, the extensive computational demands of iGPT, primarily due to the quadratic complexity of attention mechanisms relative to sequence length, limit its applicability for various vision tasks. For iGPT, this sequence length corresponds to the total number of pixels $Seq = H \times W$. As such, iGPT is primarily suited for low-resolution images (*e.g.*, $Seq = 32 \times 32$).

To mitigate this computational challenge, especially for high-resolution image training, approaches like SAIM (Qi et al., 2023) and RandSac (Hua et al., 2022b) have been developed. A critical advancement in these methodologies is the incorporation of Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020), which significantly transforms the tokenization approach — instead of treating each pixel as an individual token, ViT redefines tokens as image patches (*e.g.*, clusters of pixels). This strategy effectively reduces the sequence length for each image, thereby enabling the practical application of iGPT to higher-resolution images.

3.2. D-iGPT

Our development of D-iGPT is built upon iGPT with the ViT architecture. Unlike iGPT completely drops the knowledge of the 2D input structure, D-iGPT is designed to carefully encode this information. Specifically, at the input level, images are divided into multiple equally-sized, non-overlapping regions, forming clusters $S = \{s_1, ..., s_n\}$. Note that each cluster contains multiple spatially neighbored image patches, and serves as a fundamental unit in the sequence for autoregressive modeling. This encoding is crucial for facilitating a more intricate interplay between different regions (rather than "local" patches) of an image, thereby enhancing the effectiveness of autoregressive modeling.

Consequently, the autoregressive probability, previously defined for individual pixels in iGPT (as in Equation 3), is now reformulated for these clusters as:

$$p(s) = \prod_{i=1}^{n} p(s_i | s_1, ..., s_{i-1}, \Theta) \tag{5}$$

By default, we configure the number of clusters to 4, corresponding to a dimension of 112×112 for an input image of 224×224 (*e.g.*, each cluster contains 7×7 images patches of the size 16×16), as illustrated in Figure 2.

Building upon this new setup, we next introduce two simple yet essential modifications to enhance iGPT.

Modification I: semantic tokens. In contrast to the inherently semantically-rich nature of text, raw pixels in images generally lack such depth of meaning. Addressing this semantic discrepancy is crucial for enhancing learning efficacy

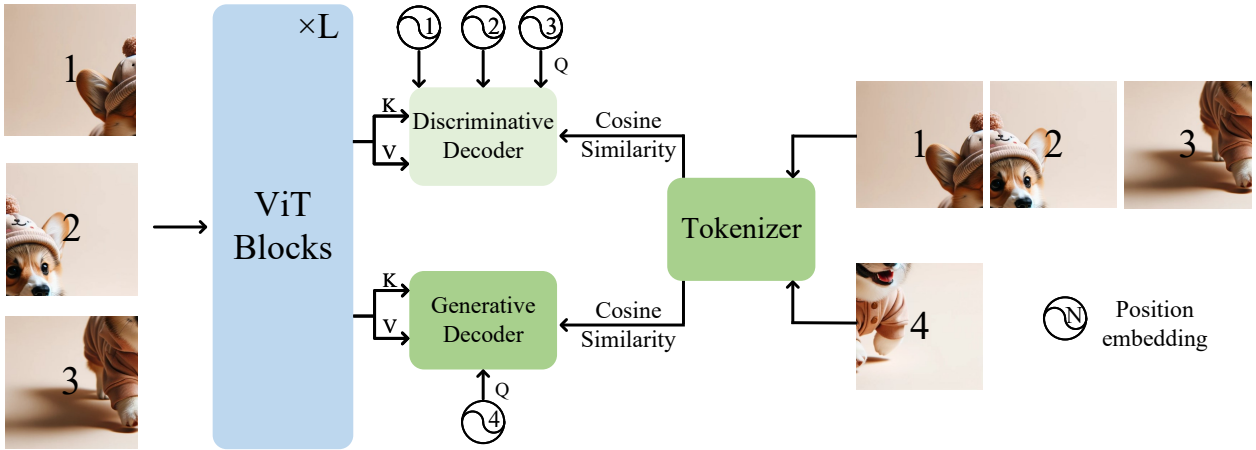


Figure 2. The overview illustration of D-iGPT.

in models like iGPT. To bridge this gap, our approach, inspired by BEiT (Bao et al., 2022), involves transitioning the autoregressive target of D-iGPT from raw pixels to semantic tokens, which can be written as:

$$\mathcal{L}_G = -\sum_{i=1}^{n} cosine(G(f(x_{s_1:s_{i-1}}); \theta_G), f_\phi(x)_{s_i}), \quad (6)$$

where $f(\cdot)$ is the encoder, $f_\phi(x)_{s_i}$ is the semantically enriched tokens corresponding to the cluster s_i, $G(\cdot;\ \theta_G)$ is the generative decoder for autoregressive prediction, and $cosine$ is the cosine similarity loss.

Furthermore, to break the dependency on a fixed sequence order and enhance learning flexibility, we adopt strategies from (Hua et al., 2022a; Yang et al., 2019) by randomly permuting the sequence of clusters $\{s_1, ...s_n\}$ and selecting a permutation π.

Mofication II: supervision on visible clusters. To further enhance the training of our model, we introduce additional supervision targeting visible clusters. This is formulated as:

$$\mathcal{L}_D = -\sum_{i=1}^{n} cosine(D(f(x_{s_1:s_{i-1}}); \theta_D), f_\phi(x)_{s_1:s_{i-1}}) \quad (7)$$

where $D(\cdot;\ \theta_D)$ is the discriminative decoder, tasked with predicting the semantic tokens of visible pixels.

This approach, as encapsulated in Equation (7), can be conceptualized as a form of knowledge distillation (Hinton et al., 2015b) — its objective is to enable the encoder of D-iGPT (the student model) to distill knowledge from the model $f_\phi(x)$ (the teacher model), which provides semantic tokens, based on the visible sequence of clusters $\{s_1, ..., s_{i-1}\}$. However, our methodology differs from traditional knowledge distillation frameworks (Wei et al., 2022; Hinton et al., 2015a), which typically align logits or feature maps between teacher and student models directly. Instead, we apply the knowledge distillation supervision on the separately designed discriminative decoder $D(\cdot;\ \theta_D)$. This design helps to disentangle different supervisions in training (*i.e.*, autoregressive on $G(\cdot;\ \theta_G)$, distillation on $D(\cdot;\ \theta_D)$), and is crucial for ensuring the acquisition of high-quality representations, as demonstrated in the subsequent experimental section.

Summary. The integration of these two modifications significantly enhances the capabilities of iGPT for visual representation learning. While there are various options for $f_\phi(x)$ to generate semantic tokens, our empirical findings, as detailed next, indicate a marked preference for discriminatively trained models like CLIP (Radford et al., 2021).

Moreover, from an implementation perspective, we adopt the attention mask strategy as employed in (Radford & Narasimhan, 2018; Chen et al., 2020a; Hua et al., 2022b; OpenAI, 2023). This approach facilitates efficient computation of input sequences of varying lengths (*e.g.*, a set of input sequences such as $\Big\{\{s_1\}, \{s_1, s_2\}, ..., \{s_1, s_2, ..., s_{n-1}\}\Big\}$) within a single iteration. We direct interested readers to the supplementary materials for more details.

Architecture design. The D-iGPT architecture is composed of two primary components: the encoder and the lightweight decoders. For the encoder, it leverages the standard ViT architecture. For the lightweight decoders, each incorporates two Transformer decoder blocks by default. Note that while the discriminative decoder D and the generative decoder G share the same architecture design, they are characterized by different sets of parameters. As shown in Figure 2, they take the randomly initialized [Dis] tokens D or [Gen] tokens G with position information as the query, and the output features from the encoder as the key and the value. Notably, in downstream tasks, we utilize only the encoder, discarding the decoder component.

Method	Pretraining Epochs	Tokenizer/Teacher	Classification	Segmentation
Base-size models (ViT-B)				
DeiT (Touvron et al., 2020)	300	Label	81.2	47.2
BEiT (Bao et al., 2022)	800	DALLE	83.2	-
MAE (He et al., 2022)	1600	Pixel	83.6	48.1
SdAE (Chen et al., 2022)	300	EMA	84.1	48.6
PeCo (Dong et al., 2021)	300	VQGAN	84.1	46.7
TinyMIM (Ren et al., 2023b)	300	MAE	85.0	52.2
FD (Wei et al., 2022)	300	CLIP	84.8	-
BEiTv2 (Peng et al., 2022)	300	CLIP+VQGAN	85.0	52.7
Randsac (Hua et al., 2022b)	1600	Pixel	83.7	-
SAIM (Qi et al., 2023)	800	Pixel	83.9	-
PeCo (Dong et al., 2021)	800	VQGAN	84.5	48.5
data2vec (Baevski et al., 2022)	800	EMA	84.2	-
SIM (Tao et al., 2022)	1600	EMA	83.8	-
iBOT (Zhou et al., 2021)	1600	EMA	84.0	-
MaskFeat (Wei et al., 2021)	1600	HOG	84.0	-
BEiTv2 (Peng et al., 2022)	1600	CLIP+VQGAN	85.5	53.1
DeepMIM (Ren et al., 2023a)	1600	CLIP	85.6	53.1
MILAN (Hou et al., 2022)	1600	CLIP	85.6	-
EVA (Fang et al., 2022)	800	CLIP	85.5	53.3
D-iGPT (Ours)	300	CLIP	**86.2**	**53.8**
Large-size models (ViT-L)				
BEiTv2 (Peng et al., 2022)	300	CLIP+VQGAN	86.6	55.0
BEiT (Bao et al., 2022)	800	DALLE	85.2	
MAE (He et al., 2022)	1600	Pixel	85.9	53.6
PeCo (Dong et al., 2021)	800	VQGAN	86.5	-
iBOT (Zhou et al., 2021)	1600	EMA	84.8	-
MaskFeat (Wei et al., 2021)	1600	HOG	85.7	-
BEiTv2 (Peng et al., 2022)	1600	CLIP+VQGAN	87.3	56.7
MILAN (Hou et al., 2022)	1600	CLIP	86.8†	-
D-iGPT (Ours)	300	CLIP	**87.8**	**57.3**

Table 1. Fine-tuning results which methods were pretrained on **ImageNet-1K** and fine-tuned on ImageNet-1K/ADE20K on classification and semantic segmentation. †: reproduced result using official code.

4. Experiment

Implementation details. In our experiments, we use CLIP to provide semantic tokens. We pretrain, by default, all models on ImageNet-1K dataset for 300 epochs. We set the batch size to 4096 and the peak learning rate to $lr = 1.5e^{-4} \times batchsize/256$. We adopt a cosine learning rate decay schedule with a warm-up period of 40 epochs, and utilize the AdamW (Loshchilov & Hutter, 2019) optimizer with a weight decay of 0.05. We use random resized cropping and random horizontal flipping, with the input size set to 224×224.

When further scaling the pretraining to ImageNet-21K dataset, all models undergo 150 epochs of pretraining with a warm-up stage of 5 epochs, a learning rate $lr = 1.5e^{-3}$, and a batch size of 4096.

4.1. ImageNet-1K Pretraining

For a fair comparison with previous work (Bao et al., 2022; Peng et al., 2022; He et al., 2022; Wei et al., 2021; Baevski et al., 2022; Ren et al., 2023a; Dong et al., 2021; Ren et al., 2023b), we first study pretraining on ImageNet-1K (Russakovsky et al., 2015) dataset with ViT-B and ViT-L.

4.1.1. ImageNet Classification

Following (He et al., 2022), we finetune pretrained models using the ImageNet-1K training set, and test on the ImageNet-1K validation set with the input size of 224×224.

Note that different from previous approaches such as (Zhai et al., 2022; Yu et al., 2022), which employs multi-head attention pooling, and BEiT-3 (Wang et al., 2022), which exploits an additional pretrained giant language tower as the

Method	IN-1K ↑	IN-V2 ↑	IN-Real ↑	IN-A.↑	IN-Ren.↑	IN-C. ↓	IN-S.↑	IN-H. ↑
Base-size models (ViT-B)								
DeiT (Touvron et al., 2020)	81.2	70.6	86.7	27.9	45.4	36.8	32.3	23.8
TinyMIM (Ren et al., 2023b)	85.0	75.3	88.7	43.0	54.6	32.7	41.0	29.2
MAE (He et al., 2022)	83.6	72.9	88.1	33.6	50.0	37.8	36.4	25.5
BEiT (Bao et al., 2022)	83.2	71.8	87.9	32.8	49.6	38.7	35.1	25.1
iBOT (Zhou et al., 2021)	84.0	73.0	88.2	33.0	51.2	36.9	38.7	26.3
BEiTv2 (Peng et al., 2022)	85.5	76.2	89.2	54.0	61.7	30.9	45.9	30.2
D-iGPT (Ours)	**86.2**	**76.4**	**89.6**	**56.3**	**64.3**	**29.9**	**48.5**	**31.1**
Large-size models (ViT-L)								
MAE (He et al., 2022)	85.9	76.5	89.4	56.3	61.0	31.1	45.6	32.4
BEiT (Bao et al., 2022)	85.2	75.1	88.8	55.4	59.8	32.0	43.8	31.2
iBOT (Zhou et al., 2021)	84.8	74.4	87.9	53.9	57.1	34.1	42.6	30.8
BEiTv2 (Peng et al., 2022)	87.3	78.3	90.0	68.6	70.3	25.4	53.7	36.5
D-iGPT (Ours)	**87.8**	**79.6**	**90.4**	**73.0**	**80.5**	**24.7**	**60.3**	**37.6**

Table 2. Robustness and Generalization evaluation on out-of-domain datasets.

image classification task layer, we hereby opt for a simple linear layer for classification. We finetune the pretrained model for 100 epochs.

Results. As shown in Table 1, our ViT-B impressively achieves 86.2% top-1 accuracy. This is the first instance of a ViT-B model surpassing the 86% accuracy threshold on ImageNet-1K, using an input size of 224×224.

In terms of comparative performance, D-iGPT demonstrates a significant improvement over various existing methods. It exceeds the baseline supervised model, DeiT, by a substantial margin of **+5.0%**, the prevalent mask image modeling method, MAE, by **+2.6%**, and the prior art MILAN/DeepMIM by **+0.6%**. Furthermore, with the same teacher model, D-iGPT surpasses EVA by **+0.7%**, while requiring only 37.5% of the training epochs.

When enlarging the model size to ViT-L, our D-iGPT sets a new benchmark with an accuracy of 87.8%. Notably, this result surpasses the well-known mask image modeling MAE by **+1.9%** and prior art BEiT-v2 by **+0.5%**.

4.1.2. Semantic Segmentation

For semantic segmentation, we evaluate D-iGPT using the ADE20K dataset (Zhou et al., 2019), which comprises 150 categories with 20,000 training images and 2,000 validation images. Following MAE (He et al., 2022), we adopt our D-iGPT pretrained ViT model as the backbone and UperNet (Xiao et al., 2018) as the framework. The input image resolution is 512×512 for training and evaluation; we report mIoU as the evaluation metric.

The last column in Table 1 reports the performance of D-iGPT on ADE20K. We note that D-iGPT achieves a mIOU of 53.8 with ViT-B and a mIOU of 57.3 with ViT-L, which sets new benchmarks for their respective model sizes. These impressive results highlight the strong generalization capabilities of D-iGPT on downstream tasks.

Additionally, we assess model robustness on out-of-domain samples. We note that D-iGPT consistently outperforms both supervised models like DeiT and self-supervised models like MAE across all out-of-domain datasets. We refer interested readers to our appendix for more details.

4.1.3. Robustness

We assess model robustness on various out-of-domain ImageNet datasets, including natural adversarial examples (ImageNet-A (Hendrycks et al., 2021b)), semantic shifts (ImageNet-R (Hendrycks et al., 2021a)), common image corruptions (ImageNet-C (Hendrycks & Dietterich, 2019)), image sketches (ImageNet-S (Wang et al., 2019)), ImageNet-V2 (Recht et al., 2019), ImageNet-Real (Beyer et al., 2020), and ImageNet-Hard (Taesiri et al., 2023).

As indicated in Table 2, D-iGPT consistently outperforms both supervised models like DeiT and self-supervised models like MAE across all datasets, showcasing notable improvements in robustness and generalization. For example, compared with the prior art BEiT-v2, D-iGPT exhibits superior robustness with improvements ranging from 0.2% to 2.6% in the ViT-B model size category. These improvements are even more striking with the ViT-L model, *i.e.*, D-iGPT makes significant strides in challenging datasets like IN-Adversarial (improvement of +4.4%), IN-Sketch (+6.6%), and IN-Rendition (+10.2%).

4.2. Pretraining with Larger Datasets

Next, we explore the impact of pretraining on ImageNet-21K with ~14 million samples. Following (Fang et al., 2022;

Method	Model	Model Size	Pretraining Data Category	Pretraining Data Size	ImageNet-1K top-1 (%)
TokenLearner (Ryoo et al., 2021)	TokenLearner	460M	I	300M (Private)	88.9
MaxViT (Tu et al., 2022)	MaxViT	475M	I	300M (Private)	89.5
SwinV2 (Liu et al., 2022b)	SwinV2	3B	I	84M (Private)	90.2
CoAtNet-7 (Dai et al., 2021)	CoAtNet	2.44B	I	300M (Private)	90.9
Lion (Chen et al., 2023)	ViT	2.44B	I	3B (Private)	91.1
BEiT (Bao et al., 2022)	ViT	306M	I	14M	88.6
iBOT (Zhou et al., 2021)	ViT	306M	I	14M	87.8
OpenClip-H (Cherti et al., 2023)	ViT	632M	I-T	2B	88.5
EVA (Fang et al., 2022)	ViT	1B	I	30M	89.6
BEiT (Bao et al., 2022)	ViT	1.9B	I-T,T,I	21M,160G,14M	89.5
One-Peace (Wang et al., 2023)	Transformer	4B	I-T,A-T	2B,8k hours	89.8
D-iGPT-L (ours)	ViT	306M	I	14M	89.5
D-iGPT-H (ours)	ViT	632M	I	14M	90.0

Table 3. Summary of D-iGPT on various vision benchmarks. I, T, and A indicate images, texts, and audios respectively. Method indicate using private training data.

Bao et al., 2022), we initially undertake supervised finetuning on the ImageNet-21K training dataset for 60 epochs; subsequently, we fully finetune models on the ImageNet-1K training dataset.

Main Results. The scaling results of D-iGPT, as depicted in Table 3, are particularly noteworthy. When pretrained with ImageNet-21K, D-iGPT successfully helps ViT-L to secure a top-1 accuracy of 89.5%. This performance not only parallels other baselines such as BEiT-3 and EVA, but also is attained with a considerably smaller model and training data size. By scaling the training to the larger ViT-H, we observe a further improvement, achieving an accuracy of 90.0%. This result is particularly noteworthy as it beats all existing solutions that build on public datasets; moreover, this 90.0% accuracy is even comparable to those achieved by substantially larger models that have been trained with extensive private datasets (Liu et al., 2022a; Dai et al., 2021; Chen et al., 2023). These results demonstrate the scalability and efficacy of D-iGPT for visual representation learning.

Linear probing. Following (Peng et al., 2022; El-Nouby et al., 2024), we also study model performance under the linear probing setup, *i.e.*, we freeze the backbone and linearly evaluate the performance on ImageNet-1K. Specifically, we consider different model sizes, by scaling the vanilla ViT from Base size to Huge size, and different dataset sizes, by scaling from ImageNet-1K to ImageNet-21K.

As shown in Figure 3, we can observe that 1) D-iGPT brings consistent improvements with bigger model sizes, and 2) larger datasets can help D-iGPT yield stronger performance. These observations demonstrate the strong scalability of D-iGPT. Additionally, it is worth mentioning that our best result is achieved by ViT-H pretrained on ImageNet-21k, with 85.9% linear probing accuracy.

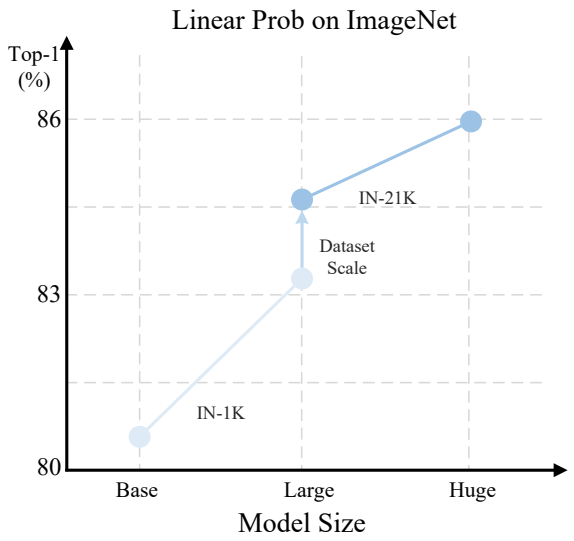


Figure 3. The model and dataset scalability of D-iGPT. D-iGPT shows significant performance gain with the growth of model size and dataset size.

4.3. Zero-shot Classification

We finetune our D-iGPT on the vision-language dataset for zero-shot ImageNet classification. With such fine-tuning, our D-iGPT can be applied to a wide array of computer vision classification tasks directly with class names, without the need for task-specific fine-tuning. Additionally, the finetuned feature can be utilized in both uni-modal and multi-modal applications (Liu et al., 2023), akin to the capabilities demonstrated by CLIP features (Radford et al., 2021).

For this process, we use the D-iGPT pretrained image en-

Pretraining	Model	DataSet	Samples	top-1
CLIPA	ViT-L/16	LAION-400M	128M	69.3
D-iGPT	ViT-L/16	LAION-400M	128M	71.6
OpenClip	ViT-L/14	LAION-400M	1B	75.3
D-iGPT	ViT-L/14	LAION-400M	1B	77.1

Table 4. Zero-shot classification performance on ImageNet-1K. Samples indicate the seen samples in finetuning.

Method	Tokenizer	Training Cost (h)	ImageNet-1K top-1 Acc.	ADE20K mIoU
MAE	Pixel	181	83.2	48.0
iGPT†	Pixel	80	82.0 $_{-1.2}$	44.1 $_{-3.9}$
MAE	VQVAE	317	83.2	47.2
iGPT†	VQVAE	100	82.3 $_{-0.9}$	47.0 $_{-0.2}$
MAE	DINO	259	84.4	50.5
D-iGPT	DINO	79	84.7 $_{+0.3}$	51.0 $_{+0.5}$
MAE	CLIP	666	85.4	52.4
D-iGPT	CLIP	159	**86.2** $_{+0.8}$	**53.8** $_{+1.4}$

Table 5. Ablation on different semantic tokens. † denotes our re-implementation with the ViT architecture. MAE is pretrained 1600 epochs while D-iGPT is pretrained 300 epochs.

coder and the OpenCLIP (Cherti et al., 2023) pretrained text encoder as our starting point. The model is then finetuned on the LAION-400M dataset (Schuhmann et al., 2021; 2022). The results, as summarized in Table4, showcase significant enhancements achieved by D-iGPT. For example, compared to CLIPA (Li et al., 2023) and OpenClip, D-iGPT improves the zero-shot ImageNet classification accuracy by 2.3% and 1.8%, respectively.

4.4. Ablation Study

Semantic tokens. Our study begins with an examination of various semantic token sources. Beyond our chosen CLIP tokens and iGPT's pixel-based tokens, we also consider alternatives like DINO features (Caron et al., 2021; Wei et al., 2021) and VQVAE tokens (Peng et al., 2022). The results, shown in Table 5, reveal notable differences in performance. While autoregressive pretraining using low-level pixels or VQVAE tokens shows lesser efficacy compared to MAE, the application of tokens from discriminatively trained models significantly enhances D-iGPT's performance, surpassing MAE by a notable margin. More importantly, with CLIP as the tokenizer, D-iGPT reduces training costs by 77% (from 666 hours to 159 hours).

Given the superior performance achieved with CLIP features, we next delve deeper into the effects of utilizing tokens from different CLIP variants. As detailed in Table 6, when we use a larger tokenizer (*i.e.*, CLIP-L), D-iGPT achieves better performance compared to using a smaller tokenizer (*i.e.*, CLIP-B). However, interestingly, if we employ

Student	Tokenizer Source	ImageNet-1K top-1 Acc.	ADE20K mIoU
ViT-B	CLIP-B	85.7	53.0
ViT-B	CLIP-L	85.9	53.3
ViT-B	CLIP-L@336	84.6	51.8
ViT-B	DINO-L	84.8	52.0
ViT-B	CLIP-L	85.9	53.3
ViT-B	OpenCLIP-L	85.9	53.2
ViT-B	OpenCLIP-H	86.2	53.6

Table 6. Ablation on tokenizer model.

Num of Clusters	Cluster shape	Top-1(%)
1	224×224	85.3
2	224×112	85.6
4	**112×112**	**86.2**
14	112×32	85.8
49	32×32	85.7
196	16×16	85.4

Table 7. Ablation study on the number of clusters.

CLIP-L@336 as the tokenizer while maintaining the input size of 224 × 224, the performance of D-iGPT drops significantly. We conjecture this is mainly due to a resolution mismatch during the training phase and the inference phase of CLIP-L@336.

Further experiments explore various large-size tokenizers, including DINO, CLIP, and OpenCLIP. One the one hand, we note that using OpenCLIP-L as the tokenizer, which is the same as CLIP-L in architecture but varies in training data, results in comparable performance to employing CLIP-L. Scaling to the even larger tokenizer, OpenCLIP-H, can further enhance D-iGPT's performance. On the other hand, we interestingly note that tokenizers like DINO do not yield comparatively favorable results. This may suggest that larger pretraining datasets and the inclusion of textual information are likely beneficial in generating high-quality semantic tokens for guiding D-iGPT's learning process.

Number of Clusters We configure the number of clusters from 1 to 196, corresponding to cluster shape ranging from 224×224 to 16×16, as shown in Table 7. The performance initially increases from 85.3% to a peak of 86.2% as the number of clusters increases from 1 to 4. However, further increasing the number of clusters causes the performance to gradually decline, reaching 85.4% at 196 clusters.

Pretraining paradigm In our evaluation of various pretraining paradigms, we consider Mask Image Modeling (MIM), Knowledge Distillation (KD), and our D-iGPT. To facilitate a fair comparison, especially for the MIM-based MAE model, we modify it to utilize CLIP features as the supervision target, moving away from the conventional pixel-

Method	ImageNet-1K top-1 Acc.	ADE20K mIoU
MAE† (He et al., 2022)	84.6	52.1
EVA (Fang et al., 2022)	85.0	52.6
KD (Wei et al., 2022)	85.0	52.5
D-iGPT	86.2	53.8

Table 8. Ablation on the pretraining paradigm. MAE† is our re-implementation with CLIP features as supervision targets.

Dec. Depth	Dec. Dim	ImageNet-1K top-1 Acc.	ADE20K mIoU
1	1024	85.6	52.8
2	1024	86.2	53.6
4	1024	86.0	53.2
2	512	85.8	53.0
2	768	85.9	53.3
2	1024	86.2	53.6

Table 9. Ablation on the decoder design.

based approach. The results are presented in Table 8.

The baseline pretraining methods, such as MAE, EVA, and KD, exhibit comparable performance levels in both ImageNet classification and ADE20K semantic segmentation. In contrast, our D-iGPT achieves markedly better results. For instance, while the highest performance among baseline models is 85.0% accuracy on ImageNet and 52.6 mIOU on ADE20K, D-iGPT significantly elevates these benchmarks to 86.2% accuracy on ImageNet and 53.8 mIOU on ADE20K. These findings underscore the potential of autoregressive pretraining, implemented in D-iGPT, as a more scalable and effective visual representation learner.

Decoder Design Our investigation begins with an examination of *Decoder Depth*. Our decoder design is lightweight, with the number of layers at most 4 (by default is 2). Intriguingly, this simpler decoder architecture not only significantly reduces GPU computational load but also enhances overall performance. As shown in Table 9, a 2-layer decoder outperforms a 4-layer decoder, even when maintaining the same decoder dimension of 1024. In contrast, MAE, by default, uses an 8-layer decoder for attaining the best performance.

Building on the success of the 2-layer decoder, we next turn our attention to the *Decoder Dimension (Dim)*. Through our experiments, we note that a reduction in decoder dimension results in a slight decrease in model performance. For example, by halving the decoder dimension from 1024 to 512, we observe an accuracy drop of 0.4% on ImageNet and a mIOU drop of 0.6 on ADE20K. This finding highlights the nuanced impact of decoder dimensionality on D-iGPT's effectiveness.

Method	Gen Decoder	Dis Decoder	ImageNet-1K (top-1 Acc.)
FD			84.9
D-iGPT		✓	84.7
D-iGPT	✓		85.5
D-iGPT†	✓		85.1
D-iGPT	✓	✓	86.2

Table 10. Ablation on the discriminative decoder. † indicates D-iGPT takes extra distillation

Discriminative Decoder. We ablate the discriminative decoder that predicts the semantic tokens of the visible pixels. Firstly, we check the setting where we remove both the discriminative decoder and the generative decoder, and implement feature distillation supervision (Wei et al., 2022) directly on the output feature map of the encoder. The corresponding result is reported in the second row of Table 10, showing 84.9% accuracy. This is 1.3% lower in accuracy than the default setup of D-iGPT (86.2%). Besides, if we keep the generative decoder and knowledge distillation, the performance is 0.4% and 1.2% lower than the generative decoder only and the default setup of D-iGPT, respectively. Next, if we keep only the discriminative decoder and remove the generative decoder, the accuracy will drop by 1.5% (from 86.2% to 84.7%). This outcome underscores the critical role of the discriminative decoder in maintaining the efficacy of the pretraining process in D-iGPT.

5. Conclusion

In this work, we introduce D-iGPT, an enhancement of iGPT that transitions the focus of autoregressive prediction from raw pixels to semantic tokens and supplements the supervision of visible pixels. This simple yet essential modification has led to a significant achievement: D-iGPT attains an impressive 90.0% top-1 accuracy on the ImageNet-1K dataset, a feat accomplished using solely publicly available datasets. We hope our D-iGPT can inspire more research on rethinking autoregressive pretraining for visual representation learning and bring fresh perspectives on building vision foundation models on publicly available data sources.

Impact Statement

D-iGPT showcases its extremely strong performance on ImageNet, which can motivate the community to rethink and further explore the potential of auto-regressive pretraining for visual representation learning. Furthermore, D-iGPT potentially helps to lay the foundation for an autoregressive framework capable of universal learning with different modalities, including vision, language, audio, and more.

Acknowledge

This work is supported by ONR with N00014-23-1-2641, TPU Research Cloud (TRC) program and Google Cloud Research Credits program.

References

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

Bao, H., Dong, L., Piao, S., and Wei, F. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1691–1703. PMLR, 13–18 Jul 2020a. URL http://proceedings.mlr.press/v119/chen20s.html.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *preprint arXiv:2002.05709*, 2020b.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *preprint arXiv:2003.04297*, 2020c.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., et al. Symbolic discovery of optimization algorithms. *arXiv preprint arXiv:2302.06675*, 2023.

Chen, Y., Liu, Y., Jiang, D., Zhang, X., Dai, W., Xiong, H., and Tian, Q. Sdae: Self-distillated masked autoencoder. *ArXiv*, abs/2208.00449, 2022.

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Dai, Z., Liu, H., Le, Q. V., and Tan, M. Coatnet: Marrying convolution and attention for all data sizes. *Advances in neural information processing systems*, 34:3965–3977, 2021.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., Chen, D., Wen, F., and Yu, N. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929*, 2020.

El-Nouby, A., Klein, M., Zhai, S., Bautista, M. A., Toshev, A., Shankar, V., Susskind, J. M., and Joulin, A. Scalable pre-training of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024.

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., and Gilmer, J. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021b.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015a.

Hinton, G., Vinyals, O., Dean, J., et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015b.

Hou, Z., Sun, F., Chen, Y.-K., Xie, Y., and Kung, S. Y. Milan: Masked image pretraining on language assisted representation. *ArXiv*, abs/2208.06049, 2022.

Hua, T., Tian, Y., Ren, S., Raptis, M., Zhao, H., and Sigal, L. Self-supervision through random segments with autoregressive coding (randsac). In *The Eleventh International Conference on Learning Representations*, 2022a.

Hua, T., Tian, Y., Ren, S., Raptis, M., Zhao, H., and Sigal, L. Self-supervision through random segments with autoregressive coding (randsac). In *The Eleventh International Conference on Learning Representations*, 2022b.

Li, X., Wang, Z., and Xie, C. An inverse scaling law for clip training. *arXiv preprint arXiv:2305.07017*, 2023.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009–12019, 2022b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://api.semanticscholar.org/CorpusID:257532815.

Peng, Z., Dong, L., Bao, H., Ye, Q., and Wei, F. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

Qi, Y., Yang, F., Zhu, Y., Liu, Y., Wu, L., Zhao, R., and Li, W. Exploring stochastic autoregressive image modeling for visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2074–2081, 2023.

Radford, A. and Narasimhan, K. Improving language understanding by generative pre-training. 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Ren, S., Wei, F., Albanie, S., Zhang, Z., and Hu, H. Deepmim: Deep supervision for masked image modeling. 2023a.

Ren, S., Wei, F., Zhang, Z., and Hu, H. Tinymim: An empirical study of distilling mim pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3687–3697, June 2023b.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Ryoo, M. S., Piergiovanni, A., Arnab, A., Dehghani, M., and Angelova, A. Tokenlearner: What can 8 learned tokens do for images and videos? *arXiv preprint arXiv:2106.11297*, 2021.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models.

Advances in Neural Information Processing Systems, 35: 25278–25294, 2022.

Taesiri, M. R., Nguyen, G., Habchi, S., Bezemer, C.-P., and Nguyen, A. Imagenet-hard: The hardest images remaining from a study of the power of zoom and spatial biases in image classification. 2023.

Tao, C., Zhu, X., Huang, G., Qiao, Y., Wang, X., and Dai, J. Siamese image modeling for self-supervised vision representation learning. *arXiv preprint arXiv:2206.01204*, 2022.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877*, 2020.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pp. 459–479. Springer, 2022.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *ICML*, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., and Zhou, C. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Wei, C., Fan, H., Xie, S., Wu, C.-Y., Yuille, A., and Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021.

Wei, Y., Hu, H., Xie, Z., Zhang, Z., Cao, Y., Bao, J., Chen, D., and Guo, B. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

Xiao, T., Liu, Y., Zhou, B., Jiang, Y., and Sun, J. Unified perceptual parsing for scene understanding. In *ECCV*, 2018.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., and Le, Q. V. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764, 2019.

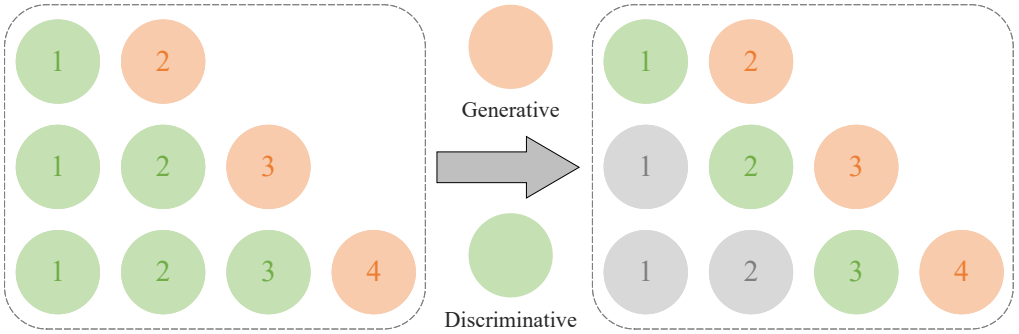
Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127 (3):302–321, 2019.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.

A. Implementation within One Trianing Iteration



We adopt the attention mask strategy in our implementation. Specifically, for an image with 4 clusters, we can have the following $\{s_1, s_2\}$, $\{s_1, s_2, s_3\}$ and $\{s_1, s_2, s_3, s_4\}$.

By default, as shown in the left part of the figure below, the supervisions applied by the Discriminative Decoder focus on $\{s_1\}$ in $\{s_1, s_2\}$, $\{s_1, s_2\}$ in $\{s_1, s_2, s_3\}$ and $\{s_1, s_2, s_3\}$ in $\{s_1, s_2, s_3, s_4\}$. We note there are redundancies in the Discriminative Decoder's supervisions, *i.e.*, in this single iteration, $\{s_1\}$ is being supervised for 3 times and $\{s_2\}$ is being supervised for 2 times.

To mitigate such redundancies, we can modify the Discriminative Decoder to supervise only on $\{s_1\}$ in $\{s_1, s_2\}$, $\{s_2\}$ in $\{s_1, s_2, s_3\}$ and $\{s_3\}$ in $\{s_1, s_2, s_3, s_4\}$ in this single iteration, as illustrated in the right part of the figure below.

To sum up, the autoregressive prediction in D-iGPT is formulated as

$$\mathcal{L}_G = -\sum_{i=1}^{n} cosine(G(f(x_{s_1:s_{i-1}}); \theta_G), f_\phi(x)_{s_i}), \tag{8}$$

where $f(\cdot)$ is the encoder, $f_\phi(x)_{s_i}$ is the semantically enriched tokens corresponding to the cluster s_i, and $G(\cdot;\ \theta_G)$ is the generative decoder for autoregressive prediction. The supervision on visible clusters is formulated as

$$\mathcal{L}_D = -\sum_{i=1}^{n} cosine(D(f(x_{s_1:s_{i-1}}); \theta_D), f_\phi(x)_{s_{i-1}}) \tag{9}$$

where $D(\cdot;\ \theta_D)$ is the discriminative decoder, tasked with predicting the semantic tokens of visible pixels.