This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Scaling Laws of Synthetic Images for Model Training ... for Now

Lijie Fan[1,†,*] Kaifeng Chen[2] Dilip Krishnan[2] Dina Katabi[1] Phillip Isola[1] Yonglong Tian[2,†]

[1]MIT CSAIL, [2]Google Research, [†]equal contribution

Github Repo: https://github.com/google-research/syn-rep-learn

Abstract

Recent significant advances in text-to-image models unlock the possibility of training vision systems using synthetic images, potentially overcoming the difficulty of collecting curated data at scale. It is unclear, however, how these models behave at scale, as more synthetic data is added to the training set. In this paper we study the scaling laws of synthetic images generated by state of the art text-to-image models, for the training of supervised models: image classifiers with label supervision, and CLIP with language supervision. We identify several factors, including text prompts, classifier-free guidance scale, and types of text-to-image models, that significantly affect scaling behavior. After tuning these factors, we observe that synthetic images demonstrate a scaling trend similar to, but slightly less effective than, real images in CLIP training, while they significantly underperform in scaling when training supervised image classifiers. Our analysis indicates that the main reason for this underperformance is the inability of off-the-shelf text-to-image models to generate certain concepts, a limitation that significantly impairs the training of image classifiers. Our findings also suggest that scaling synthetic data can be particularly effective in scenarios such as: (1) when there is a limited supply of real images for a supervised problem (e.g., fewer than 0.5 million images in ImageNet), (2) when the evaluation dataset diverges significantly from the training data, indicating the out-of-distribution scenario, or (3) when synthetic data is used in conjunction with real images, as demonstrated in the training of CLIP models.

1. Introduction

The quality and quantity of data play a crucial role in training vision models. Historically, the emphasis has been on creating large, meticulously curated image datasets with categorical labels at the image level for training supervised models [14, 34, 50, 62]. Prominent examples include CIFAR [34] and ImageNet [14]. While creating these datasets is effective on a smaller scale, their expansion to hundreds of millions of samples presents significant challenges. These challenges include the intensive labor required for curation at scale, as well as the increasing potential for noise and quality issues as the datasets scale up.

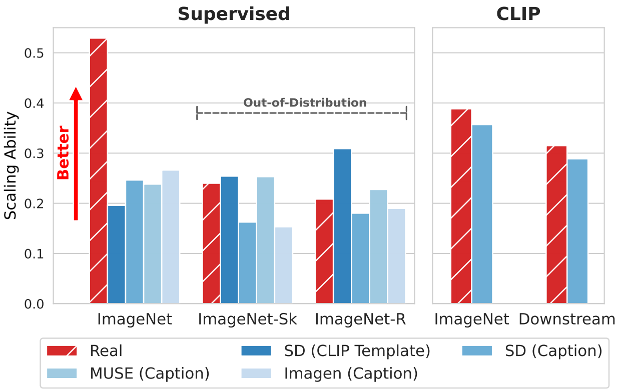

Figure 1. Scaling ability (*i.e.*, the slope of the power law curve between loss and dataset size fitted in the log space, see Eq. 2) comparison between real and synthetic images on supervised classifier and CLIP training. Red bars represent real images and blue bars represent synthetic images generated with different text-to-image models. Supervised models are trained on real or synthetic ImageNet, and text in parentheses is the text prompt used to generate the images (details in Section 3.1). ImageNet-Sketch and ImageNet-R are out-of-distribution tests. CLIP models are trained on LAION-400M with real or synthetic images. We see that: (1) scaling ability of synthetic data is *slightly worse* than that of real data for CLIP training; (2) robustness on ImageNet-Sketch and ImageNet-R datasets can be *better* when training on synthetic data.

Recently, there has been an increasing interest in training vision models using language supervision [30, 45]. This shift is exemplified by models like CLIP [45], which move beyond the fixed, predefined categories typical of datasets like ImageNet. Training these models requires extensive image-text pair datasets. Developments ranging from the creation of the Conceptual Captions dataset [59], which comprises millions of image-text pairs, to the LAION

[*] Work done while at Google.

dataset [58], encompassing billions of pairs, are examples of this growing trend. However, this approach is not without its challenges. The massive scale of data sourcing, often through web scraping, introduces significant noise. Scalability issues also persist. Moreover, the immense size of these datasets presents practical difficulties in terms of storage and data transfer. For instance, LAION-2B requires tens of terabytes of disk space and could take days, if not weeks, to download.

Fortunately, recent breakthroughs in text-to-image models have introduced exciting new possibilities in the realm of synthetic data generation. These models, capable of producing high-quality images from textual descriptions, offer several significant advantages. Firstly, they allow precise control over image content through input texts, which could provide categorical labels or paired text supervision for free. Secondly, they are bandwidth-efficient, as only the model needs to be transferred, not the entire dataset. For instance, models like Stable Diffusion [51] occupy merely 5 GB of disk space, which is 2000× more efficient compared to the massive LAION-2B dataset. Thirdly, they facilitate easier scalability with markedly reduced human labor for dataset curation. These benefits naturally lead to the question of whether it's feasible to scale up vision datasets with synthetic images for training supervised models.

However, the use of synthetic images is also not without its drawbacks. When scaled to tens or hundreds of millions of images, these models may produce images of lower quality or with misaligned concepts, and might also struggle with maintaining diversity. In this paper, we tackle a pivotal question: *How effective is the scaling of synthetic images, specifically generated for training supervised vision models?* We examine the scaling behavior of synthetic images created by cutting-edge text-to-image models, comparing their efficacy to real images in two key scenarios: the training of supervised classifiers and the training of vision models with language supervision, such as CLIP. Additionally, we explore a range of factors that markedly impact the scaling efficiency of synthetic images. These include the choice of text-to-image model, the classifier-free guidance scale employed, and the nature of text prompts used for generating training images. A summarized comparison of the scaling ability between real and synthetic images is shown in Figure 1.

We present our key findings as follows:

- An empirical study on the scaling behavior of images synthesized by three major text-to-image models (Stable Diffusion [51], Imagen [56], and Muse [9]) shows that model performance exhibits power law scaling [32] as a function of the number of synthetic images they are trained on. This trend holds until computation budget and model size become limiting factors [32].
- We identify several factors that can significantly alter the scaling ability of synthetic data, including prompt design, classifier free guidance, and the choice of models.
- In supervised settings, synthetic data does not scale as effectively as real data. However, there are exceptions where synthetic data demonstrates better scaling: (1) with classes that text-to-image models are particularly adept at generating, and (2) when the test data deviates significantly from the training data, *e.g.*, out of distribution data.
- In CLIP training, the disparity in scaling performance between synthetic and real data is less pronounced. Incorporating synthetic data with real data leads to enhanced zero-shot performance in most scenarios.

2. Related Work

Text to image models. Recent breakthroughs in text-to-image models, primarily driven by advances in diffusion models [27, 60, 71], have enabled the generation of high-quality, photo-realistic images using neural networks. Key examples of such models include Imagen [56], which performs diffusion in pixel space, and Stable Diffusion [51], which operates in the latent space of an autoencoder. DALL-E 3 [6] also exemplifies this category. An alternative family of models, based on visual tokens, utilizes VQGAN [66] and Transformers [67]. Prominent examples within this category include Parti [74] and Muse [9]. Additionally, recent advancements have been exploring the scaling Generative Adversarial Networks (GANs) [16] for text-to-image generation, as demonstrated in works such as [31].

Learning from synthetic data. Synthetic data has proven to be effective in improving performance across various domains [12, 19, 35, 38, 40, 53, 54, 63, 65, 72]. Synthetic images, in particular, have been extensively utilized in a range of different computer vision tasks, including object detection [44, 55], semantic segmentation [10, 52], autonomous driving [1], and robotics [41, 73]. More recently, there has been evidence that combining synthetic images generated by text-to-image models with real images can improve the performance on supervised learning tasks [18]. Particularly, [4, 75] have fine-tuned the text-to-image model using the target dataset, *e.g.* ImageNet, while this paper studies the capabilities of off-the-shelf text-to-image models. Additionally, there are efforts developing methods for learning transferable representations from synthetic images [5, 18, 29, 37, 49, 57, 64].

Neural scaling laws. Scaling up model size, data amount, and training budget has unlocked new capabilities of deep models [11, 13, 42, 46, 76]. Recent studies [24, 32] suggest the testing loss behaves as a power low with respect to each of these three resources when the other two are proper, in large language models (LLMs), machine translation [17], auto-regressive generative models [22], and transfer learning [23]. Similar behavior is observed in multi-modal models [2]. Chinchilla [28] suggests scaling up data proportionally to model size, to obtain compute-optimal LLMs. [3] propose to fit scaling laws by extrapolating training curves. Of particular interest, [61] theoretically shows one can break the power law with respect to data size with an ideal data pruning strategy. In this paper, we focus on the scaling behavior of synthetic data for training models.

3. Preliminaries

We first study the scaling behavior of synthetic images generated with state-of-the-art text-to-image models under the ImageNet supervised training setting.

3.1. Three Factors on T2I Generation

There are three primary factors influencing the generated images used for supervised training: (1) choice of text-to-image model, (2) the classifier-free guidance scale, and (3) the class-specific prompt used for the text input. We will now provide a detailed description of each of these factors:

Text-to-Image Models. We conducted the study on three state-of-the-art text-to-image models of different types:

- *Stable Diffusion [51]*, a model that drives the diffusion process in the latent space of a pre-trained autoencoder.
- *Imagen [56]*, a model that drives the diffusion process directly in the raw pixel space.
- *Muse [9]*, a visual token-based generation model trained with masked generative modeling, that performs discrete diffusion in the latent space of an autoencoder.

These models have distinct architectural designs, but are all capable of generating photo-realistic images. Since Imagen [56] and Muse [9] are not publicly available, we base our work on a version trained on internal data sources.

Guidance Scale. All modern text-to-image models primarily rely on the classifier-free guidance (CFG) technique to generate images based on textual input [26]. Increasing the CFG scale typically improves the alignment between the generated images and the input text, resulting in higher-quality output images. However, this also tends to reduce the diversity of content in the generated images. Through empirical analysis, we determined that when generating images for training supervised classifiers, it is advisable to use a relatively ***lower*** CFG scale compared to the default value used in generation. This ensures that the generated images exhibit a higher degree of diversity, particularly when generating images from texts describing the same class. We conducted a detailed analysis and determined the optimal CFG scale ranges for different models: [1.5, 10.0] for Stable Diffusion, [1.0, 2.0] for Imagen, and [0.1, 1.0] for Muse.

Class-specific Prompts. To generate images for each class in ImageNet, we employed different techniques to create corresponding text prompts. This allows us to generate images conditioned on the specific ImageNet class via the prompts. Take the class 'Tench' as example, we can have prompts as:

- *Classnames:* Directly use the ImageNet class name. ('Tench')
- *Classnames + Description:* Combine class name with its WordNet [39] description. ('tench, freshwater dace-like game fish of Europe and western Asia ...')
- *Classnames + Hypernyms:* Combine ImageNet class name with its Wordnet hypernyms. ('Tench, Tinca tinca, cyprinid, cyprinid fish')
- *Word2Sen:* Use a pre-trained T5 model [47] as used in [18] to convert the ImageNet class name into a sentence. We generate 100 sentences for each class. ('a tench with fish in the distance.')
- *CLIP templates:* Generate either 7 or 80 sentences with the text templates CLIP used for zero-shot classification task. ('a photo of the large tench')
- *IN-Captions:* Combine the class name with captions from ImageNet(IN) training images. Captions are generated by BLIP2 [36]. ('Tench, a man holding a fish')

3.2. Metrics: Recognizability and Diversity

The above factors give us a number of configurations to generate synthetic data. We now proceed to define metrics to analyze the resulting images, and then analyze the scaling behavior exhibited by the images generated under this configuration. The generated images should possess two crucial attributes: (1) Recognizabilty: Synthetic images should exhibit high precision, meaning they correctly represent the intended class, and high recall, implying that images for other classes should not mistakenly contain elements of this class. (2) Diversity: It is essential that the generated images are diverse from each other to improve generalization.

We define two measures to quantify the recognizability and diversity of images generated under a specific configuration. We generate 50 images for each ImageNet class, resulting in a synthetic test set comprising 50,000 images. Subsequently, we define the two metrics as follows:

- *Recognizabiliy:* Use a pre-trained ImageNet classifier (a ViT-B with 86.2% accuracy from [69]) to classify the generated images and compute the F1 score for each class. The final metric is given by averaging F1 score across all classes.
- *Diversity*[1]*:* Following [8], we extract features from the same pre-trained model [69] and compute the standard deviation on the feature space for images from every class, and then compute the average score across all classes.

3.3. Scaling Law for Synthetic Data

Prior works on scaling laws, such as [32], have observed that, for *sufficiently large* models, the test loss L and dataset

[1] We also tried replacing the diversity metric with FID [25] or LPIPS [77], please refer to Appendix D for details.

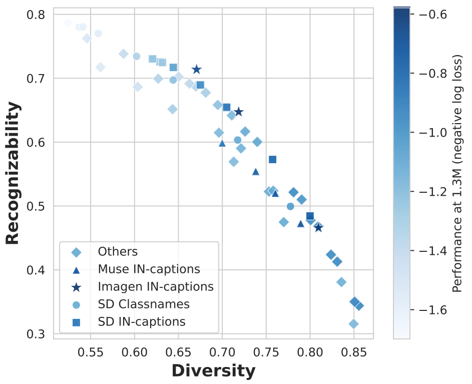

Figure 2. Recognizability vs. diversity plot for various synthetic image generation configurations (as in Section 4.2), colored by the performance at 1.3M on ImageNet validation set (measured by negative log loss). Deeper color stands for smaller loss and better performance.

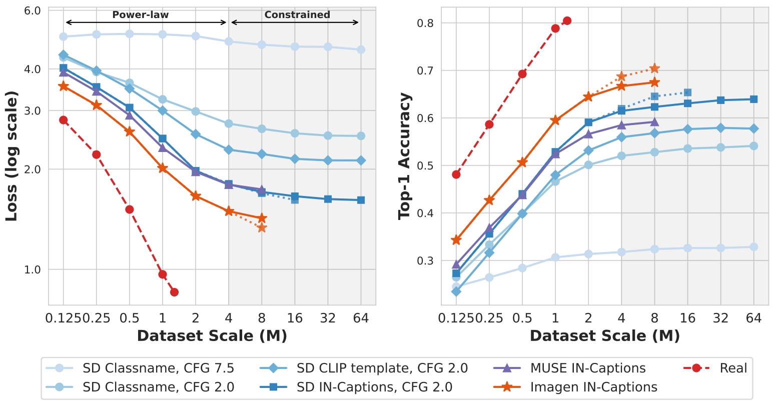

Figure 3. Scaling on ImageNet validation set for various configurations as in Section 4.3. Loss and data scale follows the power-law (as in Equation 2) with varied k when data is less than 4M. By tuning the CFG scale, text prompts and text-to-image models, the scaling behavior for synthetic images can be significantly improved (from light blue to orange). Red dashed line is for real images. Orange and blue dotted lines are ViT-L backbones, extending the power-law to 8M.

scale D, approximately follow a power-law relationship:

$$L_D \propto (1/D)^k \tag{1}$$

where k is a constant. Thus L_D exhibits linear dependence on D in log space. Let D_I be 1.3 million, roughly the size of the ImageNet training set with real data. We re-write Equation 1 as:

$$\log L_D = \underbrace{-k}_{k\text{: Scaling Ability}} (\log D - \log D_I) - \underbrace{(-\log L_{D_I})}_{\text{Performance at 1.3M}} \tag{2}$$

The slope $-k$ and y-intercept $-\log L_{D_I}$ would determine a unique scaling curve in log space. With this, we provide quantitative definitions for two key metrics for scaling:

- **Scaling Ability**: Quantifies the scaling effectiveness of synthetic images generated by a particular text-to-image configuration. Stepper curves means loss scales better with data, therefore we represent scaling ability by the negative of the **slope**: k.
- **Performance at 1.3M**: Measures the classification performance of models (as negative log loss) when trained on a dataset with a scale equivalent to 1.3M, the size of the ImageNet training set. It is represented by the **y-intercept** $-\log L_{D_I}$.

4. Scaling on Supervised Training

4.1. Setup

We train supervised classification models exclusively using the images generated by text-to-image models and evaluate their performance by computing cross-entropy loss and top-1 accuracy on the ImageNet validation set, which contains real images. Training iterations are scheduled linearly based on the training data size in logarithmic space. All generated images are resized to a resolution of 256x256 pixels. Unless stated otherwise, we employ the ViT-B model [15] with a patch size of 16 as our backbone architecture. Training hyperparameters details are provided in Appendix A.

4.2. Performance at 1.3M

We commenced by generating synthetic ImageNet datasets, each containing 1.3 million synthetic images, using various configurations of text-to-image models, CFG scales, and prompts as outlined in Section 3. In total, we created synthetic ImageNets in 54 distinct configurations, with detailed information provided in Appendix C. Figure 2 displays the validation loss on the real ImageNet validation set, represented by $-\log L_{D_I}$ as defined in Equation 2. A higher value correlates with increased classification accuracy, signaling better performance. Comparisons focusing on classification accuracy are also included in Appendix C.

Within this study, we investigate the impact of different prompt sets (Section 3) within the Stable Diffusion model. Squares (■), Circles (●), and Diamonds (◆) represent prompt configurations involving IN-captions, Classnames, and all other prompt setups, respectively. For Muse and Imagen configurations, we maintain the prompt set as IN-Captions and vary the CFG scale within the ranges [1, 2] and [0.1, 1], respectively. Triangles (▲) represent the performance of images generated with Muse, while Stars (★) represent the performance of images generated with Imagen. Several key findings emerge from the results:

Diversity and Recognizability trade-off: Across different configurations, we observe a trade-off between diver-

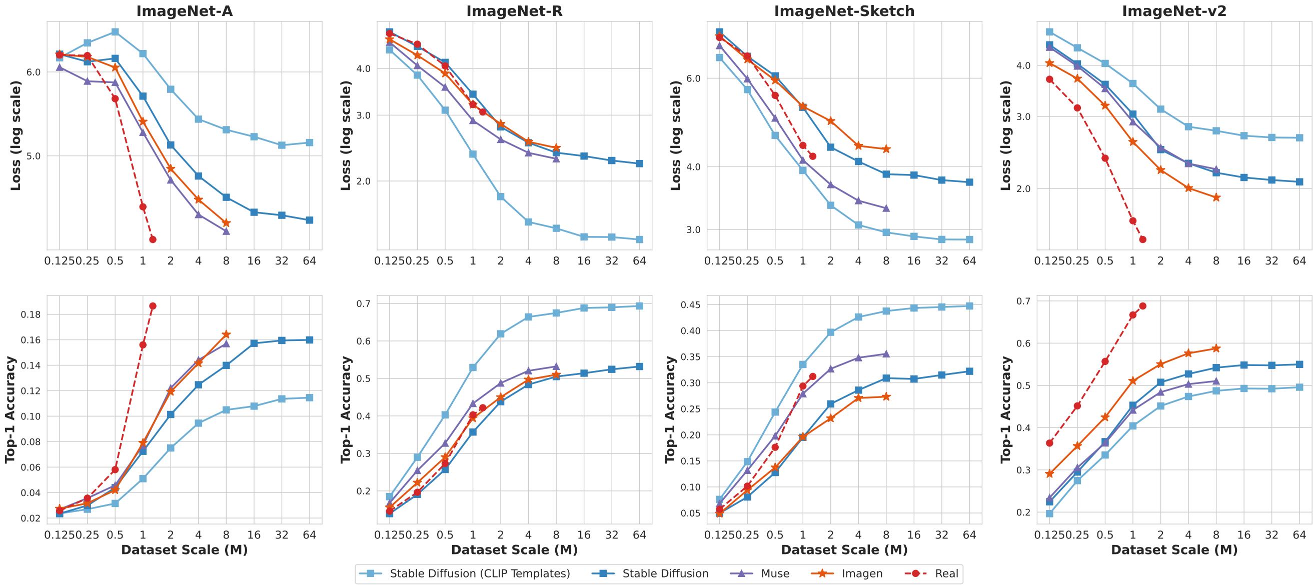

Figure 4. Scaling behavior on four different out-of-distribution validation sets. We compare synthetic images generated with optimal CFG scales by Stable Diffusion (with 80 CLIP templates or IN-Captions prompt), Imagen and Muse (all with IN-Caption prompt) with real images. Scaling synthetic data is useful and can surpass real images when the domain gap between the training and testing is significant, e.g. when evaluated on ImageNet-R and ImageNet-Sketch.

sity and recognizability. The top-right corner of the figure represents the best performance, indicating configurations that can generate both accurate and diverse images. Configurations perform poorly when either recognizability or diversity falls below a certain threshold. Also see Figure A4 bottom left for scattering colorized by accuracy.

Effect of Prompt Sets: Choosing different prompt sets can impact performance. Using a more diverse prompt set shifts the configuration towards the bottom-right of the figure. Transitioning from Classname to IN-captions for text-to-image prompts may contribute to this shift, likely due to the increased diversity on the text side, which inherently leads to more diverse generated images.

Impact of CFG Scale: When prompts are fixed, controlling the CFG scale also affects the performance of the classification model. Increasing the CFG scale shifts the configuration towards the upper-left part of the figure, where recognizability is increased, but diversity decreases. This initially leads to improved performance, followed by a decrease.

Text-to-Image Model Performance: In terms of text-to-image models, when all configured to use IN-Captions as prompts, Stable Diffusion, Imagen, and Muse demonstrate a quite similar trend in balancing recognizability and diversity. This similarity in their trade-off is reflected in their close proximity to each other in the plot.

4.3. Scaling Ability

We next proceed to analyze the scaling behavior of different models, as well as the difference between training supervised models on synthetic images and on the real ImageNet training set. Figure 3 illustrates the scaling behavior across various configurations. Specifically, for Stable Diffusion, we depict the scaling behavior of different configurations with various prompts and CFG scales. We select the optimal configuration for Muse and Imagen from Section 4.2, using IN-Caption as prompts and the corresponding optimal CFG scale for each model. From the figure, several observations can be made:

Power-law Relationship: Training on synthetic images follows a power-law relationship from 0.125 million to 4 million training images. Validation loss and training data size exhibit a linear correlation when analyzed in log space.

Scaling Disparity: Training on synthetic images does not scale as effectively as training on the real ImageNet training set images, and typically has a smaller scaling ability. This difference can be attributed to the curation of ImageNet training images and performing validation under an in-domain setting.

Impact of Prompts and CFG Scale: Using default prompt sets and CFG scale for image generation results in poor scaling ability, i.e. a very flat slope and smaller k value. However, by tuning the prompts and CFG scale properly, the generated images become much more diverse, leading to an increased scaling behavior for synthetic images, bringing it closer to the scaling ability observed with real images. Nevertheless, the best scaling configuration is still significantly worse than scaling with real data.

4.4. Scaling beyond 4M

We naturally wonder about the scaling behavior when the dataset size exceeds 4 million images and whether the validation loss will continue to decrease. In Figure 3, we also illustrate the scaling curve for Stable Diffusion up to 64 million images, and for Muse and Imagen up to 8 million images (in gray background). The results indicate that the relationship changes when the dataset scale exceeds around 4 million images.

We hypothesize that this could be due to the loss being constrained by insufficient model capacity. According to [32], the power-law relationship between validation loss and training dataset size requires the model to have sufficient capacity to fit the dataset and converge. Therefore, when the dataset size exceeds 4 million images, and if we continue to use ViT-B as the backbone architecture, the validation loss in log space no longer exhibits a linear trend. To address this, we retrain the supervised models with ViT-L as the backbone architecture for the best Stable Diffusion and Imagen configuration, as shown in the dotted lines. This improvement in model capacity could achieve a lower validation loss and maintain a roughly linear ratio up to the 8 million scale and slightly postpones the inflection point.

4.5. Out-Of-Distribution Scaling

We also investigate the scaling behavior on out-of-distribution (OOD) validation sets to determine whether it differs from the in-domain setup on the ImageNet validation set. We employ the supervised ImageNet classifiers and test them on four OOD validation sets, which include ImageNet-A [21], ImageNet-R [20], ImageNet-Sketch [68], and Imagenet-V2 [48]. The scaling curves for validation loss and top-1 validation accuracy are presented in Figure 4.

Our empirical results indicate that in scenarios where the domain gap is relatively small, such as ImageNet-v2, the scaling behavior mirrors the observation in in-domain setups, with real images showing superior scaling performance. However, a more intriguing observation emerges when the domain shift is bigger, as seen in ImageNet-R and ImageNet-Sketch. In these instances, the disparity in scaling capabilities between synthetic and real images narrows. Consequently, scaling up synthetic images becomes particularly beneficial and useful. Remarkably, in situations with sufficiently large dataset scales, synthetic images can even outperform real images from ImageNet training set (e.g. for ImageNet-R and ImageNet-Sketch with Muse), highlighting the potential of synthetic images in bridging significant domain gaps. Interestingly, when images are generated with 80 CLIP templates as text prompt (the light blue line in the plot) instead of IN-Captions, the improvements over real images on ImageNet-R and ImageNet-Sketch are more significant, although the scaling ability on the ImageNet validation set is worse (as shown in Figure 3). This suggests that carefully crafting text prompts can unlock further potential in increasing the efficacy of synthetic images, particularly for OOD scenarios.

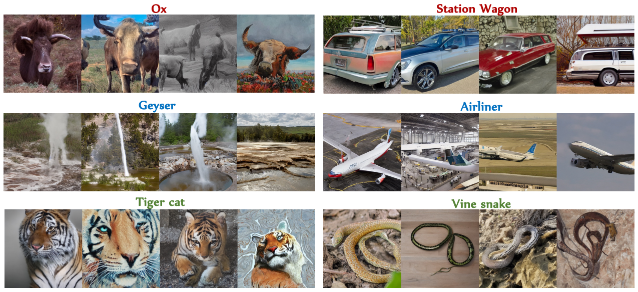

Figure 5. Visualization of different class categories generated by Stable Diffusion. Top row are the 'strong' classes that scales well. Middle row are the 'easy' classes that has a good initial performance. Bottom row are the 'poor' classes that has poor scaling ability and performance.

4.6. Zoom-in: Per Class Analysis

In addition to the general analysis of scaling behavior and its impact on overall performance, we also assess the scaling ability of each specific class in the 1,000 categories in ImageNet. For this analysis, we use on images generated by Stable Diffusion, using the optimal CFG of 2.0 and IN-Caption prompts. We created a scatter plot with scaling ability on the X-axis and 1.3M performance on the Y-axis, as shown in Figure 6. Each class is a dot, with top-right positions indicating better performance when scaled up to 4 million images. Points are colored based on either diversity or recognizability, or final performance at 4M dataset scale.

Based on their positioning in the scatter plot, classes can be categorized into three groups. Points in the bottom-left section are 'Poor' classes, which have both limited scaling ability and poor overall performance. Classes located in the upper-right section are 'Easy' ones with strong initial performance as well as robust scaling ability. Lastly, classes in the mid-right section are 'Scaling' ones. These classes may exhibit poor initial performance but demonstrate considerable improvement as the dataset size increases.

In Figure 7, we showcase two classes from each of the 'Scaling', 'Easy', and 'Poor' categories to illustrate and compare their scaling behaviors against real images. We can see certain 'Scaling' classes demonstrate a scaling ability that surpasses that of real images, emphasizing the potential utility of synthetic images in these scenarios. We present additional results for 'Scaling' classes in Appendix G.2.

Additionally, we present visualizations of the generated images from these categories in Figure 5. Our findings show that text-to-image models adeptly generate images for 'Scaling' and 'Easy' classes with commendable accuracy and diversity. However, these models face challenges in accurately rendering the correct concepts for 'Poor' classes.

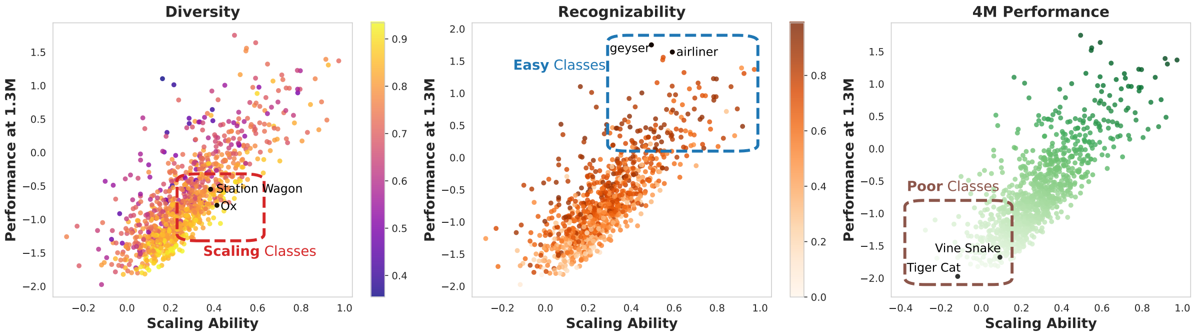

Figure 6. Scaling ability vs. Performance at 1.3M plot for synthetic data. Each point represents one of the 1,000 ImageNet classes. Classes are colored by their diversity, recognizability, and their final performance at 4M scale in the three sub-figures respectively. The scaling ability is measured by k defined in Equation 2. The performances at Y-axis is measured by the validation loss: $-\log(L)$, and higher numbers indicate lower loss and better performance. We choose two classes in each of the 'Scaling', 'Easy' and 'Poor' class categories, and their detailed scaling behavior and visualization can be found in Figure 5.

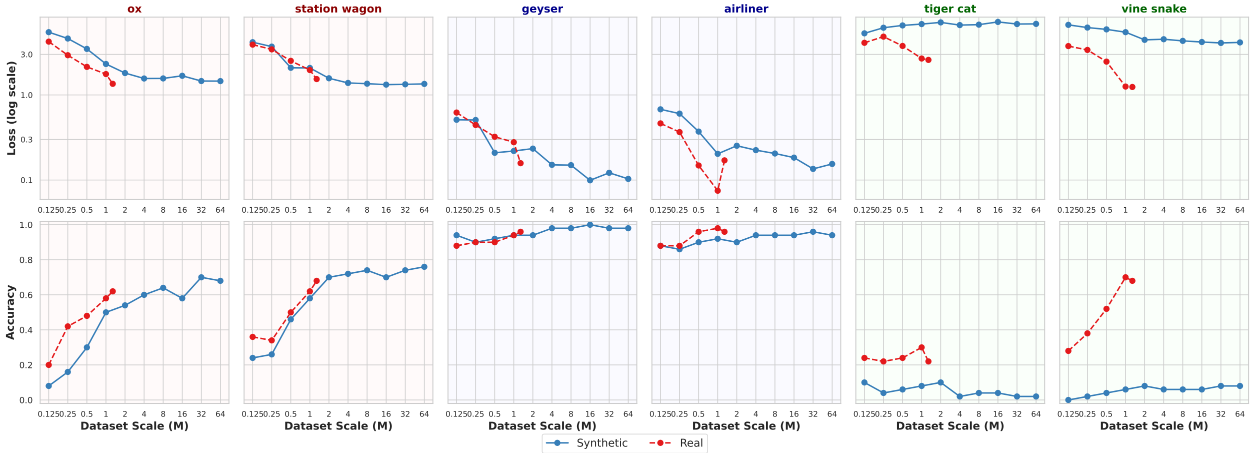

Figure 7. Scaling behavior for classes from 'Scaling (Red)', 'Easy (Blue)' and 'Poor (Green)' categories. Easy classes have a good initial accuracy with limited amounts of data, while Poor classes do not scale well. Scaling classes scale the best, and can achieve better performances than real images as the data amount goes up.

5. Scaling on CLIP

5.1. Setup

We investigated the scaling behavior of synthetic data in CLIP training using the extensive LAION-400M dataset. The synthetic images were generated using Stable Diffusion. We compare across different CFG scales and choose the optimal one (1.5) for CLIP training. For evaluation, we followed the prompt templates from [45] and conduct zero-shot classification on ImageNet and 15 different fine-grained classification datasets, including Food-101 [7], Stanford Cars [33], Oxford Pets [43] etc. The training scale begins with 1 million image-text pairs, progressively scaling up to encompass the full dataset of 371[2] million samples. All models use ViT-B as the backbone with a patch size of 16, and are trained for 32 epochs across all dataset scales. Detailed training hyper-parameters are in Appendix B. Comparisons on different CFG scales are also available in Appendix H.

5.2. Scaling Analysis

We evaluated the scaling behavior across three data setups: (1) only synthetic images, (2) only real images, and (3) a combination of both synthetic and real images. Dataset scale here refers to the number of captions. When combining synthetic and real images for training, we maintained a consistent text scale throughout. During each training iteration, we randomly selected one image, either real or synthetic, for use. The comparative analysis of these setups, evaluated on zero-shot classification loss and accuracy on ImageNet validation set, is depicted in Figure 8.

The analysis revealed that for all three scenarios, zero-shot classification loss adheres to the power-law relationship when the data amount is under around 64 million, com-

[2] The LAION-400M dataset we used contains slightly less samples compared to the orignal one because of link rot.

Table 1. Zero-shot transfer performance on 15 downstream datasets. Models are trained on LAION-400M subsets at various scales from 1M to the total 371M, with images from synthetic, real or synthetic+real. Combining synthetic images with real images can improve performance, especially when data amount is limited.

Scale	Data	Food-101	CIFAR-10	CIFAR-100	SUN397	Cars	Aircraft	DTD	Pets	Caltech-101	Flowers	STL-10	EuroSAT	RESISC45	GTSRB	Country211	**Average**
1M	Syn	5.2	12.8	3.3	5.9	1.7	0.9	5.5	6.7	17.8	3.5	29.4	9.0	9.7	5.4	1.2	7.9
	Real	5.2	25.4	7.6	5.0	2.1	1.0	5.4	5.4	18.0	5.0	36.4	14.7	9.3	6.6	1.0	9.9
	Syn+Real	**10.9**	**32.2**	**13.0**	**13.1**	**4.6**	**1.4**	**9.4**	**12.0**	**36.0**	**8.9**	**62.5**	**19.8**	**14.7**	**7.5**	**1.9**	**16.5**
4M	Syn	19.6	19.7	7.1	23.0	22.4	2.1	17.0	30.1	53.4	13.9	64.4	12.8	21.1	5.3	3.1	21.0
	Real	30.8	63.8	33.4	26.2	27.7	1.8	18.8	33.9	61.7	18.9	79.2	**40.2**	21.5	10.0	3.4	31.4
	Syn+Real	**40.3**	**67.1**	**39.3**	**35.8**	**40.9**	**2.3**	**22.9**	**45.4**	**70.1**	**23.1**	**88.2**	33.5	**27.7**	**12.3**	**4.5**	**36.9**
16M	Syn	44.2	32.4	11.5	41.6	51.3	4.9	27.4	58.3	72.1	24.8	83.6	16.7	29.5	4.6	5.9	33.9
	Real	62.9	85.2	58.1	49.0	60.6	**5.0**	30.4	61.9	81.5	**40.9**	93.1	**43.2**	39.4	**28.0**	7.4	49.8
	Syn+Real	**64.8**	**87.5**	**61.0**	**53.7**	**63.3**	**4.9**	**36.5**	**67.7**	**82.8**	**38.6**	**94.5**	**37.6**	**48.2**	**28.6**	**8.2**	**51.9**
128M	Syn	63.7	45.1	15.9	52.3	67.1	9.3	37.8	75.7	80.5	39.1	93.2	8.0	35.7	10.1	9.5	42.9
	Real	**81.9**	90.5	**70.9**	62.5	78.7	10.7	46.0	**85.9**	88.7	**60.4**	96.0	**48.3**	57.8	42.7	**14.2**	62.3
	Syn+Real	81.6	**91.0**	70.4	**64.0**	**79.4**	**11.9**	**52.5**	85.1	**90.2**	59.5	**97.0**	47.3	**61.1**	**45.3**	14.1	**63.4**
371M	Syn	70.1	51.9	26.2	55.5	70.8	12.3	41.5	79.6	83.6	45.5	95.7	28.8	39.3	20.6	10.9	48.8
	Real	**85.7**	**93.9**	**75.6**	**67.5**	**83.3**	14.2	50.1	**88.8**	91.1	**67.0**	97.0	43.9	**66.6**	42.8	**17.5**	65.7
	Syn+Real	84.6	92.4	73.2	67.1	82.0	**17.2**	**56.8**	86.4	**91.7**	61.6	**97.3**	**52.2**	65.9	**46.7**	16.0	**66.1**

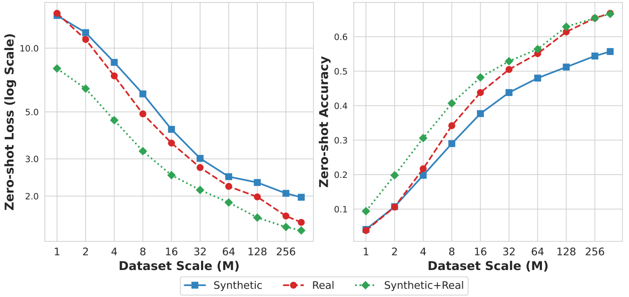

Figure 8. Scaling behavior for CLIP models trained on LAION-400M subsets of different scales. Models are trained with synthetic, real, or a combination of synthetic and real images, and are evaluated with ImageNet zero-shot accuracy. Dataset scale here refers to the number of captions.

pared to the 4 million scale in supervised training. In this range, the loss and data scale maintain a linear relationship in logarithmic space. Additionally, while the scaling efficiency (reflected in the slope of the curve) of synthetic data is lower than that of real data, this discrepancy is less pronounced than in the supervised classifier settings. However, a noticeable performance gap persists between synthetic and real images, which is likely attributable to concept mismatches between generated images and corresponding texts in certain classes, as discussed in Section 4.6.

Moreover, our results indicate that combining synthetic and real images during CLIP training can significantly enhance zero-shot performance, particularly when the dataset is limited. For instance, in training scenarios with fewer than 10 million image-text pairs, integrating synthetic images with real data can boost performance by up to 5%.

5.3. Scaling on downstream datasets

We followed the same setup and extended our comparison to include the scaling behavior of synthetic versus real images on 15 fine-grained classification datasets, detailed in Table 1. This analysis indicates a scaling behavior in these datasets that is consistent with our findings from the ImageNet evaluations. Notably, a combination of synthetic and real images demonstrated superior performance in most scenarios, particularly when the total dataset size was under 100 million samples. In cases with extremely limited data availability, such as with just 1 million samples, training on synthetic images occasionally yielded better performance than with real images, for some certain tasks, such as Pets [43] and SUN397 [70].

6. Discussion

In this paper, we investigate the scaling laws of synthetic data in model training and identify three key factors that significantly influence scaling behavior: the choice of *models*, the *classifier-free guidance scale*, and the selection of *prompts*. After optimizing these elements and increasing the scale of training data, we find that, as expected, synthetic data still does not scale as effectively as real data, particularly for supervised classification on ImageNet. This limitation largely stems from the inability of standard text-to-image models to accurately generate certain concepts. However, our study also highlights several scenarios where synthetic data proves advantageous: (1) In certain classes, synthetic data demonstrates better scaling behavior compared to real data; (2) Synthetic data is particularly effective when real data is scarce, for instance, in CLIP training with limited datasets; (3) Models trained on synthetic data may exhibit superior generalization to out-of-distribution data.

References

[1] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *IJCV*, 2018. 2

[2] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. *arXiv preprint arXiv:2301.03728*, 2023. 2

[3] Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. *NeurIPS*, 2022. 3

[4] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[5] Manel Baradad Jurjo, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In *NeurIPS*, 2021. 2

[6] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions, 2023. 2

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 7

[8] Victor Boutin, Thomas Fel, Lakshya Singhal, Rishav Mukherji, Akash Nagaraj, Julien Colin, and Thomas Serre. Diffusion models as artists: Are we closing the gap between humans and machines? *arXiv preprint arXiv:2301.11722*, 2023. 3

[9] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2, 3

[10] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 2

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[12] Yabo Dan, Yong Zhao, Xiang Li, Shaobo Li, Ming Hu, and Jianjun Hu. Generative adversarial networks (gan) based efficient sampling of chemical composition space for inverse design of inorganic materials. *NPJ Computational Materials*, 2020. 2

[13] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[15] 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. *arXiv preprint arXiv:2010.11929*, 2020. 4

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 2

[17] Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *EMNLP*, 2021. 2

[18] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*, 2022. 2, 3

[19] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Nlp with synthetic text. *TACL*, 2022. 2

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021. 6

[21] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 6

[22] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 2

[23] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021. 2

[24] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 2

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 3

[26] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[28] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[29] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 2

[30] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[31] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, 2023. 2

[32] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2, 3, 6

[33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013. 7

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[35] Varun Kumar, Ashutosh Choudhary, and Eunah Cho. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*, 2020. 2

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 3

[37] Hao Liu, Tom Zahavy, Volodymyr Mnih, and Satinder Singh. Palm up: Playing in the latent manifold for unsupervised pretraining. *arXiv preprint arXiv:2210.10913*, 2022. 2

[38] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *arXiv preprint arXiv:2202.04538*, 2022. 2

[39] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 3

[40] Masato Mimura, Sei Ueno, Hirofumi Inaguma, Shinsuke Sakai, and Tatsuya Kawahara. Leveraging sequence-to-sequence speech synthesis for enhancing acoustic-to-word speech recognition. In *SLT*, 2018. 2

[41] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *CoRL*, 2022. 2

[42] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[43] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 7, 8

[44] Xingchao Peng, Baochen Sun, Karim Ali, and Kate Saenko. Learning deep object detectors from 3d models. In *CVPR*, 2015. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 7

[46] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 2

[47] 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. *JMLR*, 2020. 3

[48] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 6

[49] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In *CVPR*, 2018. 2

[50] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 1

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[52] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2

[53] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech recognition with augmented synthesized speech. In *ASRU*, 2019. 2

[54] Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP*, 2020. 2

[55] Artem Rozantsev, Vincent Lepetit, and Pascal Fua. On rendering synthetic images for training an object detector. *CVIU*, 2015. 2

[56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 2, 3

[57] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, 2023. 2

[58] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2

[59] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1

[60] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[61] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *NeurIPS*, 2022. 3

[62] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 1

[63] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models.*, 2023. 2

[64] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv preprint arXiv:2306.00984*, 2023. 2

[65] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 2020. 2

[66] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 2

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[68] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 6

[69] Ross Wightman. Pytorch image models. `https://github.com/huggingface/pytorch-image-models`, 2019. 3

[70] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 8

[71] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 2

[72] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. *arXiv preprint arXiv:2004.11546*, 2020. 2

[73] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Tsung-Yi Lin, Alberto Rodriguez, and Phillip Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *ICRA*, 2022. 2

[74] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 2

[75] Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. *arXiv preprint arXiv:2310.10402*, 2023. 2

[76] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022. 2

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3