Fake it till you make it: Learning transferable representations from synthetic ImageNet clones

Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, Yannis Kalantidis

Fake it till you make it: Learning transferable representations from synthetic ImageNet clones

Mert Bulent Sariyildiz[1,2] Karteek Alahari[2] Diane Larlus[1] Yannis Kalantidis[1]

[1] NAVER LABS Europe [2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK

Abstract

Recent image generation models such as Stable Diffusion have exhibited an impressive ability to generate fairly realistic images starting from a simple text prompt. Could such models render real images obsolete for training image prediction models? In this paper, we answer part of this provocative question by investigating the need for real images when training models for ImageNet classification. Provided only with the class names that have been used to build the dataset, we explore the ability of Stable Diffusion to generate synthetic clones of ImageNet and measure how useful these are for training classification models from scratch. We show that with minimal and class-agnostic prompt engineering, ImageNet clones are able to close a large part of the gap between models produced by synthetic images and models trained with real images, for the several standard classification benchmarks that we consider in this study. More importantly, we show that models trained on synthetic images exhibit strong generalization properties and perform on par with models trained on real data for transfer. Project page: https://europe.naverlabs.com/imagenet-sd

1. Introduction

The rise of (shallow) machine learning [15, 85] and later deep learning [27, 46, 80] has entirely changed the landscape of computer vision research over the past few decades, shifting some of the focus from *methods* to the *training data* itself. Datasets, initially of hundreds of images and dozens of classes [22, 23], have grown in size and complexity, and started becoming contributions in their own right. They have been fueling the progress of computer vision as much as, if not more than, the methods themselves. ImageNet [17], and mainly its ImageNet-1K [71] subset of about 1 million annotated images, has impacted the field in an unprecedented way. Yet, curating and annotating such a dataset comes at a very high money and labor cost.

The last couple of years have seen the rise of large and generic models, trained on data which is less curated but

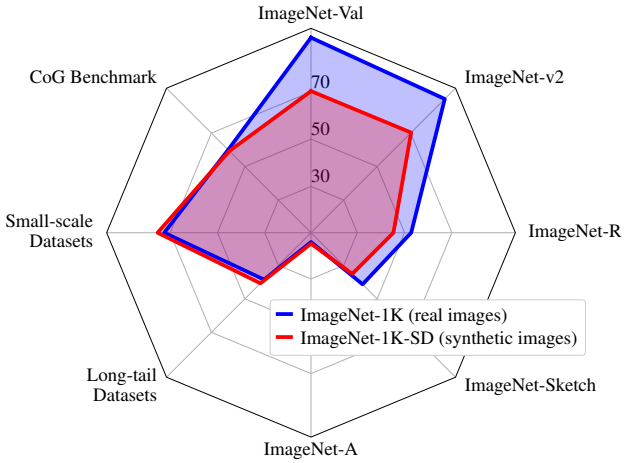


Figure 1. **ImageNet-1K vs ImageNet-1K-SD.** The blue polygon shows the performance of a model trained on ImageNet-1K. The red polygon depicts the performance of one trained on ImageNet-1K-SD, *i.e.*, only on synthetic data generated with Stable Diffusion [70] using the class names of ImageNet-1K. We report top-5 accuracy for ImageNet test sets, and average top-1 for transfer tasks.

orders of magnitude larger. Those proved to be easily applicable, either directly, or combined with a tailored model, to a wide range of computer vision transfer tasks [38, 42, 65]. They have also been used beyond prediction tasks, *e.g.*, for text-conditioned image generation. Models such as DALL-E [66] or Stable Diffusion [70] have demonstrated impressive image generation ability. They produce fairly realistic synthetic images and exhibit a high degree of compositionality.

Such generative models are trained on billion-scale datasets [76] composed of noisy image-text pairs scraped from the internet. Although training such models is out of reach for most institutions, a few of them have been made available to the community. Given the remarkable ability of these generative models, it is only natural to ask provocative questions such as: *Is there still a need for real images when training image prediction models?*

In this paper we explore this question through one of the most iconic computer vision datasets, ImageNet [17]. We study to which extent this dataset can be entirely replaced

by synthetic images when learning deep models. For this, we assume that we are provided with a set of classes, and the Stable Diffusion [70] model a generator that can produce realistic images from a textual prompt.

Our task is to learn an image classification model *from scratch* using a dataset composed only of synthetic images. We then evaluate the performance of this model on several datasets. First and foremost, we measure how well models and classifiers trained only on synthetic images recognize the training classes in real images from the standard ImageNet validation set. Then, we evaluate them on common datasets that test their resilience to domain shifts or adversarial examples, still for the ImageNet training classes. Finally, we consider several transfer learning scenarios where we measure the generalization performance of our models to novel classes. Fig. 1 summarizes the main results by comparing models trained on two equally sized set of images from the same set of classes, one real and one synthetic, on a number of these tasks. The gap is surprisingly narrow, especially for some of these scenarios.

To summarize, our contributions are threefold. First, we leverage Stable Diffusion [70] and generate synthetic ImageNet clones, *i.e*., datasets with synthetic images for the ImageNet classes, using class names as prompts. We analyse the generated images, highlight important issues, and propose class-agnostic alterations to the basic prompt that reduce semantic issues and increase diversity. Second, we train classification models using different ImageNet clones and show that they can achieve 91.7% and 70.3% top-5 accuracy on ImageNet-100 and ImageNet-1K respectively. Finally, we evaluate the generalization capacity of our models. We show that their performance gap with models trained on real images is reduced when testing for resilience to domain shifts or adversarial examples. Moreover, we show that our models perform on par with models trained conventionally when testing on 15 transfer datasets.

2. Related work

2.1. Learning with synthetic data

Learning with synthetic data has become a standard way to create large amounts of labeled data for annotation heavy tasks, such as human understanding [64, 84], semantic segmentation [14, 73], optical flow estimation [19, 90] or dense visual alignment [63]. In most cases, this synthetic data requires access to 3D models and renderers [53], or to a simulator [69] with a physically plausible engine. Recent works propose pretraining on a database of synthetic fractal [43] or sinusoidal wave [81] images before fine-tuning the model using real images on a downstream task. In this study we use synthetic data to learn encoders and classifiers that can be used *out-of-the-box*, without the need for a subsequent fine-tuning step. Closest to our work, Kumar *et al*. [78] generate synthetic OCT images to train a glaucoma detection model to be applied to real images. Here, we target synthetic clones of complex natural image datasets, *i.e*., ImageNet-1K [71], and we use a *general-purpose* text-to-image generation model.

Synthetic ImageNet clones. Synthetic images for ImageNet classes have been used recently in a number of related works [2, 48, 67] based on class conditional Generative Adversarial Networks (GANs), such as BigGAN [6]. Besnier *et al*. [2] generate images for ten ImageNet classes and propose techniques to reduce the gap between models trained on generated images and real ones. Li *et al*. [48] synthesize five images for each ImageNet-1K class, together with their semantic segmentation annotations to automatically generate pixel-level labels at scale. Our work focuses on image-level classification, and uses a general-purpose text-conditioned generative model instead of ImageNet-1K class-conditioned GANs. It further offers a larger scale study with promising results on the full ImageNet-1K benchmark when training from 1.28 million synthetic images. Concurrent work [28] also synthesizes data for ImageNet-1K, but focuses on improvements on top of the CLIP [65] model or after fine-tuning.

Synthetic images as data++. Data sampled from generative models [25, 33, 66, 70] can be seen as data with added functionalities or "data++" [39]. Such data can be manipulated, interpolated or composed [11, 12, 40, 41] with dedicated operators in their latent space, and further used for counterfactual reasoning [49, 55, 59]. In this paper, we do not exploit these added functionalities. Our prompts consider a class at a time and do not leverage any interpolation nor the composition properties of synthetic data. Instead, we chose our complete pipeline, including the set of data augmentations, to be identical to the one we use for real images, to allow for a fair comparison.

Zero-shot learning and test-time view synthesis. Generative models have been used to extend models to new classes, or to create novel views at test time. Chai *et al*. [12] synthesize novel views for test-time ensembling by perturbing the latent code of a test image. Aiming at zero-shot recognition [92], Elhoseiny *et al*. [21] synthesize a classifier for any novel class given its semantic description (*e.g*., textual or attribute-based), whereas others synthesize images [20, 26], or image *features* [47, 74] using such descriptions. Here we aim to learn encoders from scratch, and do not rely on models previously trained on real data.

2.2. Distillation of datasets and models

Knowledge distillation [7, 32] is a mechanism to transfer knowledge from a pretrained "teacher" model into a "student" one, and it usually requires images. Our approach can be seen as performing *image-free* distillation from a generic text-to-image generation model into a specific classification model. We assume no access to images to

distill from and, instead of distilling the visual encoder of the image generation model, inspired by recent works in NLP [51], we prompt a generation model to produce synthetic images and train a classifier with them.

Dataset distillation [10, 96], on the other hand, is a way of compressing a training set of real images into a smaller set of synthetic images such that after training a model on those, it performs as well as if it had been trained on the original set. However, one needs to tailor the generation process to a specific task, whereas in our case, we sample images from a task-agnostic generator.

Reconstructing images from model activations can be considered as another form of distillation. Earlier works reconstruct images from gradient-based features [87, 89] or CNN activations [52]. Since then many methods have tried to uncover the training data distribution as it is stored in the weights of a model [13, 95]. Instead of trying to recover the training distribution of the teacher image generation model, we use prompting to distill its knowledge for a specific image classification task.

3. Preliminaries

In this section, we first define the task we solve, *i.e.*, learning an image classification model when the training set of real images is replaced by an image generator, and training proceeds using only synthetically generated images. We then briefly describe Stable Diffusion [70], *i.e.*, the text-to-image generation model we use in this paper.

Task formulation. Our goal is to learn an image classification model given a set of class names $\mathcal{C}$ and a text-to-image generator $\mathcal{G}$. This task is a variant of image classification where the fixed-size image training set is replaced by an image generator. The model we aim to learn consists of an encoder $\mathbf{z} = f_\theta(\mathbf{x})$ that maps an *image* $\mathbf{x}$ into a vector representation $\mathbf{z} \in \mathbb{R}^d$, and a classifier $\mathbf{y} = q(\mathbf{z})$ that outputs a distribution $\mathbf{y}$ over the N classes $c_i \in \mathcal{C}$, where $i = \{1, .., N\}$. We follow the common supervised learning setting [46, 71] and, unless otherwise stated, learn the encoder parameters θ together with the classifier q for the task. This model (encoder and classifier) is evaluated on the initial classification task, by applying it to real images (Sec. 5.1 and Sec. 5.2). We also evaluate the visual encoder in the context of several transfer learning tasks (Sec.5.3).

Text-to-image with Stable Diffusion. We use the recent Stable Diffusion model [70] (SD) as text-to-image generator $\mathcal{G}$. SD is a denoising diffusion model [33] built around the idea of *latent diffusion*. The diffusion process is run on a compressed latent space for efficiency. An image encoder/decoder is used to interface the latent diffusion model with the pixel space. The generation process can be conditioned in many ways, *e.g.*, with text for text-to-image generation, or an image latent vector for image manipulation.

The text-to-image SD model consists of three main components: i) an autoencoder whose visual encoder outputs a structured latent representation that is fed as input to the forward diffusion process and whose decoder is then used to convert the latent vectors back to pixels, ii) a denoising U-Net that runs the diffusion process, and iii) a text encoder, *i.e.*, similar to the one used by CLIP [65].

The text-to-image generation process takes a textual prompt p as input and generates an image $\mathbf{x} \in \mathbb{R}^{W \times H \times 3}$. Let $g(p)$ denote the generation function of model $\mathcal{G}$. Image $\mathbf{x}$ is then given by $\mathbf{x} = g(p)$. In practice, the prompt p is first encoded via the text encoder and the text embedding is used as a conditioning vector for the latent diffusion process that runs for a number of steps. The latent representation is then provided to the decoder, which outputs the image $\mathbf{x}$.

There are two important parameters that control the quality and speed of text-conditioned diffusion; the number of diffusion steps and the coefficient that weights the textual conditioning vector. The former is linearly related to extraction time, while the latter provides an excellent way of controlling the visual diversity of generated images. The default values are 50 steps and guidance scale equal to 7.5.

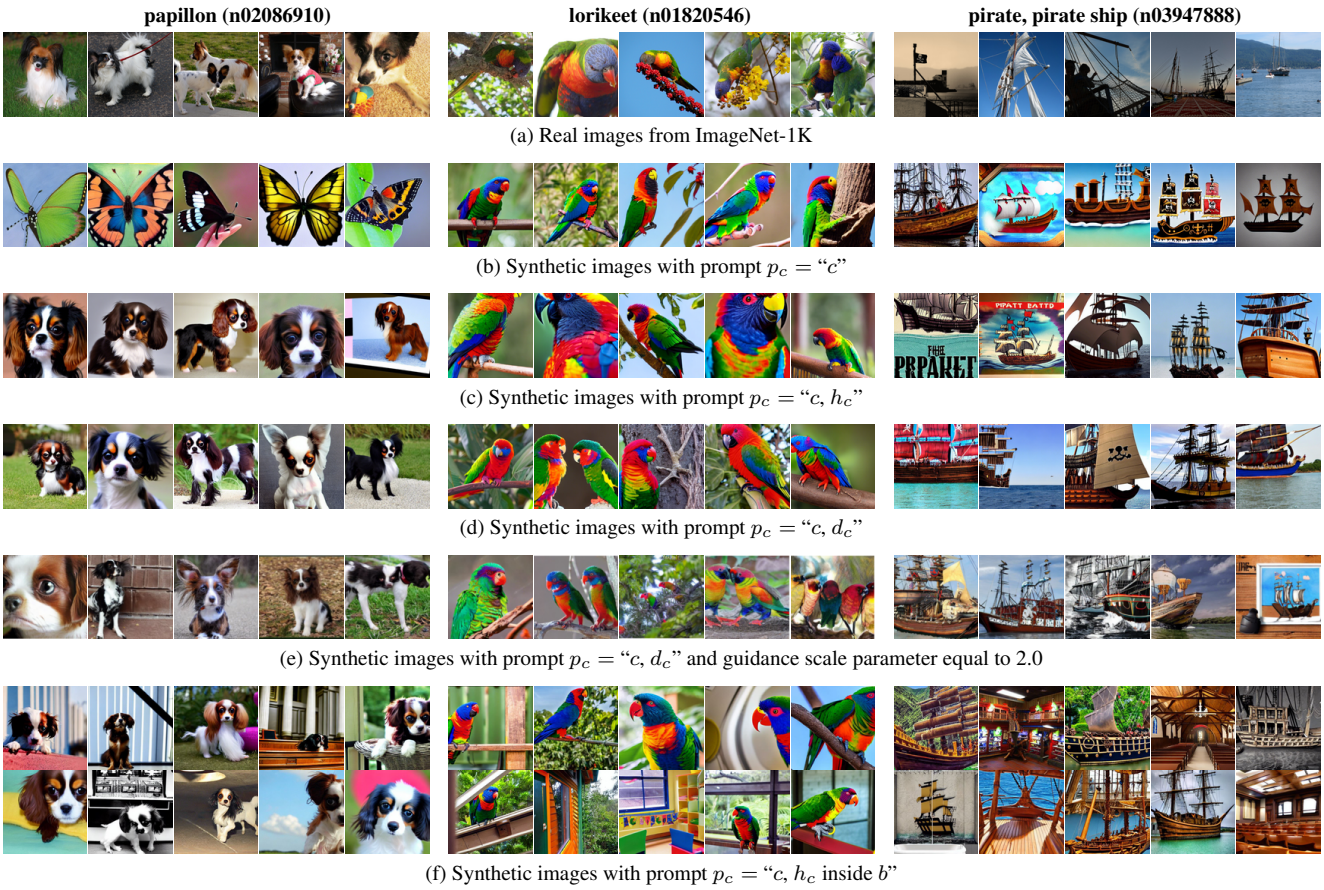
Link to distillation. Since the generator is a model that internally encodes visual information, the image classification model we learn is essentially derived from $\mathcal{G}$. Under this formulation, and as discussed in Sec. 2, one can also see this task as text-guided, image-free knowledge distillation. Here we distill knowledge from a model of a very different nature, *i.e.*, a text-to-image generation model, to a purely visual encoder, for solving a specific task.

4. Generating synthetic ImageNet clones

For our study, we create clones of the ImageNet [17] dataset by synthesizing images depicting the classes it contains. We refer to all synthetic datasets of ImageNet classes that are created using Stable Diffusion as **ImageNet-SD**. Sec. 4.1 describes different ways of creating ImageNet-SD datasets starting from simply using the class name as the prompt. We then present generic, class-agnostic ways for tackling issues that arise with respect to semantics and diversity in Secs. 4.2 and 4.3, respectively. We present a few sample qualitative results in Fig. 2, with a more extensive set in the supplementary material.

4.1. Generating datasets using class names

In the absence of a training set of real images, we use the generator $\mathcal{G}$ presented in the previous section to synthesize images for each class in the set $\mathcal{C}$. To do so, we need to provide the generator with at least one prompt per class. When used as an input, this class-conditioned prompt p_c triggers the generation of a synthetic image $\mathbf{x}_c = g(p_c)$ from class c. The simplest prompt one could think of is the class name *i.e.*, $p_c =$ "c". Although CLIP [65] uses $p_c =$ "a



(a) Real images from ImageNet-1K

(b) Synthetic images with prompt $p_c =$ "c"

(c) Synthetic images with prompt $p_c =$ "c, h_c"

(d) Synthetic images with prompt $p_c =$ "c, d_c"

(e) Synthetic images with prompt $p_c =$ "c, d_c" and guidance scale parameter equal to 2.0

(f) Synthetic images with prompt $p_c =$ "c, h_c inside b"

Figure 2. **Qualitative results.** (a) Real ImageNet images. (b)-(g) Synthetic ImageNet-SD images generated with different prompts. Despite high photo-realistic quality, some issues are noticeable for (b) such as i) semantic errors *e.g.*, for the class "papillon", ii) lack of diversity, and iii) distribution shifts *e.g.*, towards cartoons for the "pirate" class. Such issues are addressed with more expressive prompts in (c)-(g).

photo of a c" for their zero-shot experiments, using only the class name gives better results in our case.

Each class in ImageNet is associated with one or more *synsets*, *i.e.*, entities, in the WordNet [57] graph. We use the synset lemmas corresponding to each class as class-name prompt "c", comma-separated if more than one. Fig. 2b shows random examples of images generated with such prompts. At first glance, one can appreciate the ability of the generator to create photo-realistic images given only a class name. In Sec. 5, we show that one can already obtain surprisingly good image classification results by simply training a model with this synthetic dataset.

Upon close inspection of the generated images, however, some issues become apparent: a) **semantic errors**: Images generated for some classes may capture the wrong semantics (*e.g.*, see the "papillon" class in Fig. 2b), b) **lack of diversity**: Generated images tend to look alike (an issue more apparent in the supplementary material, and c) **visual domain issues**: some classes tend to shift away from natural images towards sketches or art (*e.g.*, the "pirate ship" class in Fig. 2b). We discuss and address these issues in the following.

4.2. Addressing issues with semantics and domain

As mentioned earlier, by comparing the (real) images from ImageNet with the synthetic ones generated using only synset names as prompts, we observe that for some classes their semantics do not match. This is due to polysemy, *i.e.*, multiple semantic meanings or physical instantiations of the class names we used as prompt. We show one such case in the left-most column of Fig. 2b: the "papillon" images correspond to butterfly for our generated dataset, while the ImageNet synset contains images of the dog breed of the same name (see Fig. 2a).

To reduce this semantic ambiguity, we leverage once again the fact that class names correspond to WordNet [57] synsets. We augment the prompt for class name c with two additional elements provided by WordNet: a) The *hypernyms* h_c of the synset as defined by the WordNet graph, *i.e.*, the class name(s) of the parent node(s) of this class in the graph; and b) the *definition* d_c of the synset, *i.e.*, a sentence-length description of the semantics of each synset. In both cases, we append this information to the prompt, which becomes

$p_c =$ "c, h_c" and $p_c =$ "c, d_c" for hypernyms and definition, respectively.

Qualitatively, we observed that issues regarding the semantics of the most problematic classes are fixed, and so are, to some extent, issues related to visual domain mismatch. These are also visible in Figs. 2c and 2d: appending the hypernym ($h_c =$ "toy spaniel") or the description ($d_c =$ "small slender toy spaniel with erect ears and a black-spotted brown to white coat") of the class "papillon" in the prompt produces images with the dog breed as the main subject. Appending the hypernym ($h_c =$ "ship") or the description ($d_c =$ "a ship that is manned by pirates") of the class "pirate ship" results in more natural-looking images rather than illustrations, reducing the domain shift.

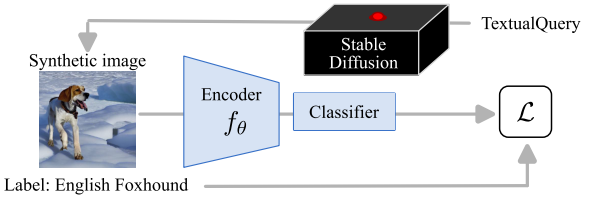
4.3. Increasing the diversity of generated images

Generating images using more expressive prompts, *e.g.*, by appending class hypernym or definition, not only reduces semantic errors, but also increases the visual diversity of the output images. This is visible, for example, in the "lorikeet" and "pirate ship" classes in Figs. 2c and 2d when compared to Fig. 2b: the pose and viewpoints are slighly more diverse. However, images still tend to display the class instance centered and in a prominent position. The real ImageNet images feature significantly more diversity, several different settings and backgrounds, and, in several cases, multiple instances of the same class (*e.g.*, see Fig. 2a).

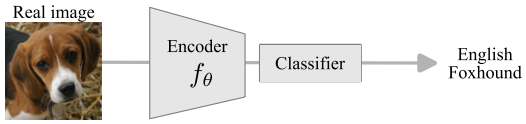
Although class-specific prompt engineering is an appealing option, in this study we chose to remain generic, and to increase diversity in class-agnostic ways.

Reducing reliance on the textual prompt. The text-conditioned generation process of Stable Diffusion uses classifier-free diffusion guidance [34] which jointly trains both the conditional and unconditional diffusion models, and combines their estimates, resulting in a trade-off between sample quality and diversity. This trade-off is controlled by the guidance scale parameter, that has in practice been shown to produce high-quality images in the range of 6-9 (the default value is 7.5). Although visually detailed (see Figs. 2b to 2d), the resulting images lack diversity. We therefore experiment with reducing the guidance scale. Despite a small degradation in the visual quality of the generated images, setting the scale to 2.0 results in more diverse sets of images as shown in Fig. 2e.

Diversifying the background. We assume that class c can be seen "inside" a scene or background. To remain class-agnostic, we use all the scene classes from the Places dataset [97] as background for every class. We generate images for every possible combination of a class c and a scene $b \in \mathcal{B}$ from the set $\mathcal{B}$ of 365 scenes in Places. We found that "c inside b" generally produces the best-looking results among a few prepositions we tried. However, we found that semantic and domain errors that arise from generating only using class name remained after specifying a background. We therefore build on top of the second simplest, but more semantically correct prompt variant, and use $p_c =$ "c, h_c inside b" to generate images in diverse scenes and backgrounds. Although we do not consider this in our study, selecting backgrounds tailored for each class, *e.g.*, by matching class names to scenes using features from a text encoder, seems like a promising future direction.



(a) Training a model on synthetic images.



(b) Testing the frozen model on real images.

Figure 3. **Overview of our experimental protocol.** During training, the model has access to synthetic images generated by the Stable Diffusion model, provided with a set of prompts per class. During evaluation, real images are classified by the frozen model.

Label noise and visual realism. Quite a few generated images, especially those with low guidance scale parameters or with random backgrounds (*e.g.*, see Figs. 2e and 2f) are not realistic, for example, the right-most image in the first column of Fig. 2e. When the prompt mentions a background, some images miss the foreground object completely (*e.g.*, see the bottom row in the middle column of Fig. 2f) or contain impossible combinations of objects and scenes. Yet, we see such noisy or unrealistic synthetic images as a way of adding stochasticity during the training process, similar to what strong non-realistic data augmentation achieves [24,94]. In fact, it was recently shown [24] that diverse data augmentations, even when inconsistent with the data distribution, can be valuable (even more than additional training data) for out-of-distribution scenarios. Our experimental validation corroborates this claim.

5. Experiments

In this section we analyze the performance of image classification models learned using the different synthetic datasets constructed as described in Sec. 4. Due to the size of ImageNet-1K (roughly 1.3 million images), we perform most of our study on the smaller ImageNet-100 [82] dataset. This allows us to run multiple flavours of each synthetic dataset and to measure the impact of several design choices. Because ImageNet-100 is a randomly chosen subset of ImageNet-1K, spanning over 100 classes and

126,689 images, it preserves some important characteristics of ImageNet-1K such as its fine-grained nature.

We denote synthetic datasets for the two ImageNet subsets as **ImageNet-100-SD** (IN100-SD) and **ImageNet-1K-SD** (IN1K-SD), respectively.

Experimental protocol. We follow the protocol illustrated in Fig. 3. The generator $\mathcal{G}$ is the Stable Diffusion [70] v1.4 model,[1] trained on the LAION2B-en dataset [76] and fine-tuned on a smaller subset filtered by an aesthetics classifier. During training, the generator is used to synthesize images for each class, which are then used for training the parameters of the encoder and the classifier. Unless otherwise stated, we create datasets of the exact same size as their real-image counterparts, *i.e.*, we generate the exact same number of images for every class as in the corresponding real dataset, maintaining any class imbalance.

We evaluate all the models on real images. When evaluating their performance over the ImageNet classes, we use both the encoder and the classifier learned during training to predict labels of real images for the 5 ImageNet datasets (Secs. 5.1 and 5.2). For transfer learning (Sec. 5.3), we use the pretrained encoder as a feature extractor, and learn a separate linear classifier on each of the 15 transfer datasets.

All our experiments use ResNet50 [27] as the encoder f_θ. Unless otherwise stated, we use 50 diffusion steps. We provide ablations for the diffusion steps and guidance scale as well as more implementation details in the supplementary material. We use multi-crop data augmentation [9], as it results in large performance gains for the models trained on ImageNet-SD (see supplementary for more details). Indeed, strong transformations have been shown to improve domain generalization [86], and to reduce the sim-to-real gap.

5.1. Results on ImageNet datasets

Evaluating different prompts on ImageNet-100. Tab. 1 compares the performance of models trained using variants of ImageNet-100-SD created with the different prompts presented in Sec. 4, for two different guidance scale values: 7.5 and 2. From the results for ImageNet-val and ImageNet-v2 (four left-most columns), we make the following observations: **(a)** Simply using the class name as a prompt and the default guidance scale (row 2), one can synthesize images and learn a visual encoder *from scratch* that already achieves *more than* 70% *Top-5 accuracy* (43% Top-1 accuracy) on ImageNet-100, a challenging 100-way classification task with many fine-grained classes. **(b)** Adding the hypernym or the definition from WordNet as part of the prompt (rows 3, 4) addresses some of the semantic and domain issues and translates into performance gains. **(c)** Generating objects on diverse backgrounds (row 5), even in a simple and class-agnostic way, gives the best results for the default guidance scale, reaching over 50% Top-1 and 76% Top-5 accuracy on ImageNet-100. **(d)** Using a lower guidance scale value (2) leads to more diverse image sets (as discussed in Sec. 4.3) and translates into the best overall performance on ImageNet-100. **(e)** The exact formulation of the prompt has less impact when lowering the guidance scale; all the four prompt variants lead to similar performance as we see from rows 6-9.

Scaling the number of synthetic images. Unlike real datasets that are capped in the number of images they contain, ImageNet-SD has theoretically no size upper bound as one can generate images on demand. We therefore generated datasets which are $10\times$, $20\times$ and $50\times$ larger than ImageNet-100, using prompt $p_c =$ "c, d_c" (the best variant in Tab. 1, row 8) for the classes of ImageNet-100. From the last three rows of the top section in Tab. 1, we see that this brings gains of up to 8.5% in Top-1 accuracy on ImageNet-100, with our best model reaching 73.3% Top-1 (and 91.7% Top-5) accuracy. The gains are even more prominent for transfer learning, as we discuss in Sec. 5.3.

Results on ImageNet-1K. In the bottom part of Tab. 1 we report results on the very challenging 1000-way classification task of ImageNet-1K (IN-Val) that contains many fine-grained categories of mushrooms, birds and dogs [36]. We see that the model trained on our synthetic ImageNet-1K-SD dataset using the prompt composed of the class name and description ($p_c =$ "c, d_c") and using guidance scale 2 reaches 42.9% Top-1 and 70.3% Top-5 accuracy on the ImageNet-1K validation set. Although significantly lower than the results achieved by a model trained on the 1.3 million real images of ImageNet, we see that the synthetic dataset is able to at least partially capture the subtle clues needed to differentiate fine-grained classes. Similar observations can be made on ImageNet-v2 [68] (IN-v2).

5.2. Resilience to domain shifts

We investigate the performance of our models on three challenging evaluation sets for ImageNet-1K classes: ImageNet-Sketch [88] (IN-Sketch), ImageNet-R [30] (IN-R) and ImageNet-A [31] (IN-A). These datasets contain out-of-distribution images and their goal is to test resilience to domain shifts and adversarial images. Results are reported in the right-most columns of Tab. 1.

For ImageNet-100, we see from the top part of the table that a number of ImageNet-100-SD models *outperform* the model trained on real images for ImageNet-Sketch and ImageNet-R. The best Imagenet-100-SD model, *i.e.* the one trained with $50\times$ images, further rivals the baseline on ImageNet-A.

When it comes to a much harder classification task like the 1000 classes of ImageNet-1K, we see from the lower part of Tab. 1 that the same trend does not really hold. The ImageNet-1K-SD model trained on synthetic data lags behind in all cases when compared to the two models [62, 91] that are trained on the ImageNet-1K training set.

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

Training Dataset	R. Size	Scale		Prompt (p_c) / *Model*	IN-Val Top-1	IN-Val Top-5	IN-v2 Top-1	IN-v2 Top-5	IN-Sketch Top-1	IN-Sketch Top-5	IN-R* Top-1	IN-R* Top-5	IN-A* Top-1	IN-A* Top-5
ImageNet-100		–	1	*Baseline*	*87.4*	*96.8*	*82.5*	*95.1*	*39.1*	*58.9*	*58.4*	*79.1*	*25.6*	*68.7*
ImageNet-100-SD		7.5	2	$p_c =$ "c"	43.1	70.7	45.4	70.7	29.9	53.5	51.7	75.3	8.8	38.4
			3	$p_c =$ "c, h_c"	46.9	73.4	47.3	73.7	25.9	50.4	46.3	75.3	11.5	42.2
			4	$p_c =$ "c, d_c"	47.9	74.2	49.1	74.9	24.7	49.2	41.2	71.5	12.2	38.5
			5	$p_c =$ "c, h_c inside b"	51.5	76.8	51.2	77.4	27.9	52.5	54.0	81.8	14.1	48.4
		2.0	6	$p_c =$ "c"	63.5	86.9	62.7	86.7	41.8	**67.6**	**64.2**	83.9	13.7	45.1
			7	$p_c =$ "c, h_c"	63.4	87.1	63.5	86.5	39.2	66.7	61.9	85.1	14.9	49.1
			8	$p_c =$ "c, d_c"	64.8	86.9	65.0	87.3	33.8	60.5	51.4	77.5	14.0	48.8
			9	$p_c =$ "c, h_c inside b"	63.1	85.7	62.0	85.0	38.7	65.5	64.0	**87.2**	**21.9**	**63.1**
	10×	2.0	10	$p_c =$ "c, d_c"	72.4	90.8	70.2	90.2	40.0	65.7	55.2	79.0	15.6	53.8
	20×		11	$p_c =$ "c, d_c"	72.4	91.4	71.4	90.7	38.4	63.9	56.9	81.5	17.8	55.0
	50×		12	$p_c =$ "c, d_c"	**73.3**	**91.7**	**72.3**	**91.2**	**42.0**	67.0	59.4	82.3	17.1	57.1
ImageNet-1K		–	13	*PyTorch [56]*	*76.1*	*92.9*	*71.1*	*90.4*	*24.1*	*41.3*	*36.2*	*52.8*	*0.0*	*14.4*
		–	14	*RSB-A1 [91]*	*80.1*	*94.5*	*75.6*	*92.0*	*29.2*	*46.5*	*40.6*	*55.1*	*11.1*	*38.6*
ImageNet-1K-SD		7.5	15	$p_c =$ "c, d_c"	26.2	51.7	26.0	51.4	9.5	22.1	15.9	32.0	2.2	10.1
		7.5	16	$p_c =$ "c, h_c inside b"	30.1	55.6	29.8	55.3	11.9	27.1	23.5	43.1	3.4	13.2
		2.0	17	$p_c =$ "c, d_c"	**42.9**	**70.3**	**43.0**	**70.3**	**16.6**	**35.1**	**26.3**	**45.3**	**3.6**	**15.1**

Table 1. **Results on ImageNet datasets.** Top-1 and Top-5 accuracy on several ImageNet datasets, namely IN-Val (the ILSVRC-2012 validation set [71]), IN-v2 [68], IN-Sketch [88], IN-R [30] and IN-A [31]. In all cases, testing is done on real images. For the prompts, h_c (d_c) refers to the hypernym (definition) of class c provided by WordNet [57], while b to scene classes from Places 365 [97]. *IN-R and IN-A only cover a subset of the ImageNet-100 classes and we compute the reported metrics only on the common classes. Brick-colored scores denote performance higher than the models trained on real images. *Italics* denote results from models trained using real images.

Training Dataset	Scale		Prompt (p_c) / *Model*	Aircraft	Cars196	DTD	EuroSAT	Flowers	Pets	Food101	SUN397	iNat18	iNat19	Avg.
–	–	1	Random Weights	11.9	3.7	17.0	73.1	26.9	11.9	13.3	7.3	0.1	1.3	16.6
ImageNet-100	–	2	*Baseline*	*43.6*	*41.5*	*67.9*	*96.2*	*85.6*	*78.7*	*63.4*	*51.2*	*22.8*	*33.4*	*58.4*
ImageNet-100-SD	2.0	3	$p_c =$ "c, d_c" (50×)	**47.9**	**44.5**	**74.0**	**96.8**	**89.6**	**83.7**	**68.6**	**57.2**	**29.5**	**40.6**	**63.2**
ImageNet-1K	–	4	*PyTorch [56]*	*48.9*	*49.9*	*72.1*	*96.2*	*89.3*	*92.3*	*71.2*	*60.5*	**35.5**	*41.5*	*65.7*
	–	5	*RSB-A1 [91]*	*46.8*	*54.4*	*73.8*	*95.8*	*88.6*	**93.0**	*71.3*	**63.4**	*34.9*	*43.2*	*66.5*
ImageNet-1K-SD	7.5	6	$p_c =$ "c, d_c"	48.7	49.7	71.6	96.5	90.1	81.9	66.4	55.8	28.7	40.6	63.0
	7.5	7	$p_c =$ "c, h_c inside b"	49.6	47.4	72.1	95.9	89.3	87.2	67.7	59.5	30.8	41.4	64.1
	2.0	8	$p_c =$ "c, d_c"	**55.3**	**57.2**	**75.9**	**96.7**	**92.9**	88.7	**73.1**	62.5	35.0	**46.3**	**68.4**

Table 2. **Top-1 accuracy on ten transfer learning datasets** for encoders trained on real and synthetic images. We treat encoders as feature extractors and train linear classifiers on top for each dataset. Brick-colored scores denote performance higher than the models trained on real images. We make the remarkable observation that representations from models trained on synthetic data can match the generalization performance of representations from models trained on millions of real images. *Italics* denote results from models trained using real images.

5.3. Transfer learning

In previous evaluations, we used pretrained models as a whole, *i.e.*, encoders together with classifiers, all trained on synthetic ImageNet datasets, and we directly applied those to predict the label of the (real) test images on the training classes. Here, we use a slightly different protocol. We evaluate the quality of the representations learned by our encoders alone, by using them as feature extractors and training linear logistic regression classifiers from scratch on top as done in transfer learning [44, 75].

We report results on 15 transfer datasets: **(a)** eight common small-scale datasets (Aircraft [54], Cars196 [45], DTD [16], EuroSAT [29], Flowers [58], Pets [61], Food101 [5], SUN397 [93]), **(b)** two long-tail datasets (iNat2018 [83] and iNat2019 [83]), and **(c)** the five datasets ("levels") of the CoG benchmark [75]. We report Top-1 accuracy on the (real) test set of the small-scale and long-tail datasets in Tab. 2. In Fig. 1 and the supplementary, we present results on the CoG benchmark. We compare ImageNet-100-SD and ImageNet-1K-SD visual encoders obtained with some of our best prompts to baselines trained

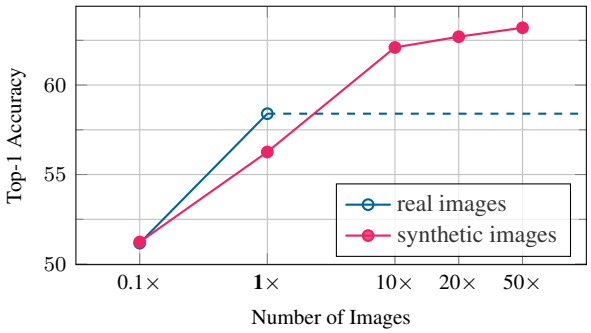


Figure 4. **Scaling the number of training images.** Average top-1 accuracy on 10 transfer datasets when training on ImageNet-100 using $(1/10)$-th to $50\times$ images (relative to the real dataset size).

on ImageNet-100 and ImageNet-1K. What we observe is quite striking: On average, representations learned on purely synthetic images exhibit *generalization performance comparable to representations trained on thousands or millions of real images*. This suggests that synthetic images can be used to pretrain strong general-purpose visual encoders.

Following this transfer learning protocol, our best model achieves *70.4%* Top-1 accuracy on ImageNet-1K (evaluation as part of the CoG benchmark, detailed in the supplementary material), significantly closing the gap to models trained on real data. This protocol differs from the one presented in Sec. 5.1 as it uses real images to train a linear classifier on top of the feature extractor trained only on synthetic images, hence results are not comparable with Tab. 1.

Scaling the number of synthetic images for transfer. Fig. 4 reports transfer learning performance on the 10 datasets of Tab. 2, when varying the size of the training set. We see that generating $10\times$ more images allows the ImageNet-100-SD model to outperform the model trained on real images, and the gains increase as we generate up to $50\times$ more.

6. Discussion

This section takes a step back and considers some of the implications from the analysis proposed in this paper.

Applicability beyond ImageNet. The process we followed to create ImageNet-SD requires minimal assumptions and can be applied to a wider set of classes. To disambiguate semantics, we only assume access to a short textual description of the class. This is generally easy to acquire even at a larger scale, *e.g.*, in semi-automatic ways from Wikipedia.

Scaling laws for synthetic data. Conceptually, there is no reason to restrict our approach to a finite dataset of synthetic images. We could devise a training process which sees each image only once [60].

Yet, despite this scaling potential, the quality of the resulting classifier is bounded by the expressivity of the generator and the concepts it can reliably reproduce. No matter how intriguing the promise of an "infinite dataset" via data generation might be, practical applications are bound by costs linked to computation and storage, as well as the moderation of the content fueling this generator. The latter has strong implications we discuss next.

Data and model bias. Because of its pioneering role as a source of images to train generic models, and all it has done to advance the computer vision field, ImageNet and some of its bias has been under heavy scrutiny [18, 50]. Its synthetic counterparts have no reason to be immune to bias.

The main advantage of training with synthetic dataset is also its biggest flaw. Instead of manually curating and annotating a dataset, this process is outsourced to a text-to-image generator, whose training data is not always known. Our study is based on the text-to-image generator of Stable Diffusion (SD). SD is trained on LAION-2B [76], a dataset scraped from the internet and filtered in an automatic way using CLIP [65]. LAION has been shown to contain problematic content [4] and SD models to memorize at least part of the training set [8, 77]. Algorithmic bias is not only due to bias in the data [35], yet biased datasets lead to biased models and predictions [1, 72, 79]. Frameworks such as [37] could be considered to increase transparency and accountability.

On top of the bias in the data, the architecture itself constraints the generated images, and as such, propagates and potentially amplifies [3] existing bias. A major one that we have discussed earlier is the lack of diversity. An obvious corollary is the fact that stereotypes are reinforced. The options we have explored mitigate this issue to some limited extent, in that it improves classification results, but this issue is far from being solved. Finally, there are many societal implications of using such models to generate synthetic datasets for training computer vision models, and a more thorough and multi-disciplinary discussion is required.

7. Conclusions

In this paper, we study to which extent ImageNet, arguably the most popular computer vision dataset, can be replaced by a dataset synthesized by a text-to-image generator. Through an extensive study, we find that one can learn models that exhibit surprisingly good performance on fine-grained classification tasks like ImageNet-100 and ImageNet-1K without any class-specific prompting. However, the most important result of this study is the finding that models trained on synthetic data exhibit exceptional generalization capability that rivals with models learned with real images. We see this study as merely a first glimpse of what is now possible with the latest large models in terms of visual representation learning. We envision that similar approaches could be used to fine-tune or adapt models, using those synthetic datasets side-by-side with real ones.

Acknowledgements. This work was supported in part by MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and the ANR grant AVENUE (ANR-18-CE23-0011).

References

[1] Osman Aka, Ken Burke, Alex Bauerle, Christina Greer, and Margaret Mitchell. Measuring model biases in the absence of ground truth. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2021. 8

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

[3] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. *arXiv:2211.03759*, 2022. 8

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963*, 2021. 8

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

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. ICLR*, 2019. 2

[7] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proc. SIGKDD*, 2006. 2

[8] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv:2301.13188*, 2023. 8

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. ICCV*, 2021. 6

[10] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proc. CVPR*, 2022. 3

[11] Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in GANs. In *Proc. ICLR*, 2021. 2

[12] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *Proc. CVPR*, 2021. 2

[13] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proc. ICCV*, 2019. 3

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

[15] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. One-class SVM for learning in image retrieval. In *Proc. ICIP*, 2001. 1

[16] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proc. CVPR*, 2014. 7

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009. 1, 3

[18] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 2021. 8

[19] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proc. ICCV*, 2015. 2

[20] Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E. Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. In *Proc. ICLR*, 2023. 2

[21] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proc. ICCV*, 2013. 2

[22] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *IJCV*, 88, 2009. 1

[23] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *Proc. CVPR-W*, 2004. 1

[24] Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? An investigation into scaling laws, invariance, and implicit regularization. In *Proc. ICLR*, 2023. 5

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11), 2020. 2

[26] Sophia Gu, Christopher Clark, and Aniruddha Kembhavi. I can't believe there's no images! learning visual tasks using only language data. *arXiv:2211.09778*, 2022. 2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1, 6

[28] 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? In *Proc. ICLR*, 2023. 2

[29] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *JSTAEORS*, 2019. 7

[30] 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. In *Proc. ICCV*, 2021. 6, 7

[31] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proc. CVPR*, 2021. 6, 7

[32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Proc. NeurIPS-W*, 2014. 2

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020. 2, 3

[34] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[35] Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2021. 8

[36] Minyoung Huh, Pulkit Agrawal, and Alexei Efros. What makes ImageNet good for transfer learning? *arXiv:1608.08614*, 2016. 6

[37] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *ACM FAccT*, 2021. 8

[38] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021. 1

[39] Phillip Isola. When faking your data actually helps – learning vision from GANs, NeRFs, and noise, 2022. BMVC Keynote talk. 2

[40] Ali Jahanian, Lucy Chai, and Phillip Isola. On the steerability of generative adversarial networks. In *Proc. ICLR*, 2020. 2

[41] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. In *Proc. ICLR*, 2022. 2

[42] 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 *Proc. ICML*, 2021. 1

[43] Hirokatsu Kataoka, Kazushige Okayasu, Asato Matsumoto, Eisuke Yamagata, Ryosuke Yamada, Nakamasa Inoue, Akio Nakamura, and Yutaka Satoh. Pre-training without natural images. *IJCV*, 2022. 2

[44] Simon Kornblith, Jonathon Shlens, and Quoc Le. Do better ImageNet models transfer better? In *Proc. CVPR*, 2019. 7

[45] Jonathan Krause, Jia Deng, Michael Stark, and Fei-Fei Li. Collecting a large-scale dataset of fine-grained cars. In *Proc. ICCV-W*, 2013. 7

[46] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NeurIPS*, 2012. 1, 3

[47] Michalis Lazarou, Tania Stathaki, and Yannis Avrithis. Tensor feature hallucination for few-shot learning. In *Proc. WACV*, 2022. 2

[48] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. BigDatasetGAN: Synthesizing ImageNet with pixel-wise annotations. In *Proc. CVPR*, 2022. 2

[49] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. In *GlobalSIP*, 2019. 2

[50] Alexandra Sasha Luccioni and David Rolnick. Bugs in the data: How ImageNet misrepresents biodiversity. *arXiv:2208.11695*, 2022. 8

[51] Xinyin Ma, Xinchao Wang, Gongfan Fang, Yongliang Shen, and Weiming Lu. Prompting to distill: Boosting data-free knowledge distillation via reinforced prompt. In *Proc. IJCAI*, 2022. 3

[52] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proc. CVPR*, 2015. 3

[53] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proc. ICCV*, 2019. 2

[54] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 7

[55] Chengzhi Mao, Augustine Cha, Amogh Gupta, Hao Wang, Junfeng Yang, and Carl Vondrick. Generative interventions for causal learning. In *Proc. CVPR*, 2021. 2

[56] Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proc. ACM-ICM*, 2010. 7

[57] George A Miller. Wordnet: A lexical database for English. *Commun. ACM*, 1995. 4, 7

[58] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proc. ICCVGIP*, 2008. 7

[59] Deniz Oktay, Carl Vondrick, and Antonio Torralba. Counterfactual image networks, 2018. 2

[60] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 2019. 8

[61] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proc. CVPR*, 2012. 7

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Proc. NeurIPS*, 2019. 6

[63] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proc. CVPR*, 2022. 2

[64] Albert Pumarola, Jordi Sanchez-Riera, Gary P. T. Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proc. ICCV*, 2019. 2

[65] 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 *Proc. ICML*, 2021. 1, 2, 3, 8

[66] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proc. ICML*, 2021. 1, 2

[67] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In *Proc. NeurIPS*, 2019. 2

[68] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proc. ICML*, 2019. 6, 7

[69] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proc. ECCV*, 2016. 2

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 1, 2, 3, 6

[71] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 1, 2, 3, 7

[72] Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv:2207.02842*, 2022. 8

[73] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proc. CVPR*, 2018. 2

[74] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. Gradient matching generative networks for zero-shot learning. In *Proc. CVPR*, 2019. 2

[75] Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *Proc. ICCV*, 2021. 7

[76] 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:2210.08402*, 2022. 1, 6, 8

[77] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. *arXiv:2212.03860*, 2022. 8

[78] Ashish Jith Sreejith Kumar, Rachel S. Chong, Jonathan G. Crowston, Jacqueline Chua, Inna Bujor, Rahat Husain, Eranga N. Vithana, Michaël J. A. Girard, Daniel S. W. Ting, Ching-Yu Cheng, Tin Aung, Alina Popa-Cherecheanu, Leopold Schmetterer, and Damon Wong. Evaluation of Generative Adversarial Networks for High-Resolution Synthetic Image Generation of Circumpapillary Optical Coherence Tomography Images for Glaucoma. *JAMA Ophthalmology*, 140(10), 2022. 2

[79] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *FAccT*, 2021. 8

[80] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015. 1

[81] Sora Takashima, Ryo Hayamizu, Nakamasa Inoue, Hirokatsu Kataoka, and Rio Yokota. Visual atoms: Pre-training vision transformers with sinusoidal waves. *arXiv:2303.01112*, 2023. 2

[82] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019. 5

[83] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proc. CVPR*, 2018. 7

[84] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proc. CVPR*, 2017. 2

[85] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *Proc. ICCV*, 2009. 1

[86] Riccardo Volpi, Diane Larlus, and Gregory Rogez. Continual adaptation of visual representations via domain randomization and meta-learning. In *Proc. CVPR*, 2021. 6

[87] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. HOGgles: Visualizing object detection features. In *Proc. ICCV*, 2013. 3

[88] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Proc. NeurIPS*, 2019. 6, 7

[89] Philippe Weinzaepfel, Hervé Jégou, and Patrick Pérez. Reconstructing an image from its local descriptors. In *Proc. CVPR*, 2011. 3

[90] Yo whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? In *NeurIPS Datasets and Benchmarks Track*, 2022. 2

[91] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. In *Proc. NeurIPS-W*, 2021. 6, 7

[92] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *PAMI*, 41(9), 2018. 2

[93] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *Proc. CVPR*, 2010. 7

[94] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *Proc. ICLR*, 2021. 5

[95] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proc. CVPR*, 2020. 3

[96] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *Proc. ICLR*, 2021. 3

[97] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017. 5, 7