Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models

Gowthami Somepalli [1], Vasu Singla [1], Micah Goldblum [2], Jonas Geiping [1], Tom Goldstein [1],

[1] University of Maryland, College Park
{gowthami, vsingla, jgeiping, tomg}@cs.umd.edu

[2] New York University
goldblum@nyu.edu

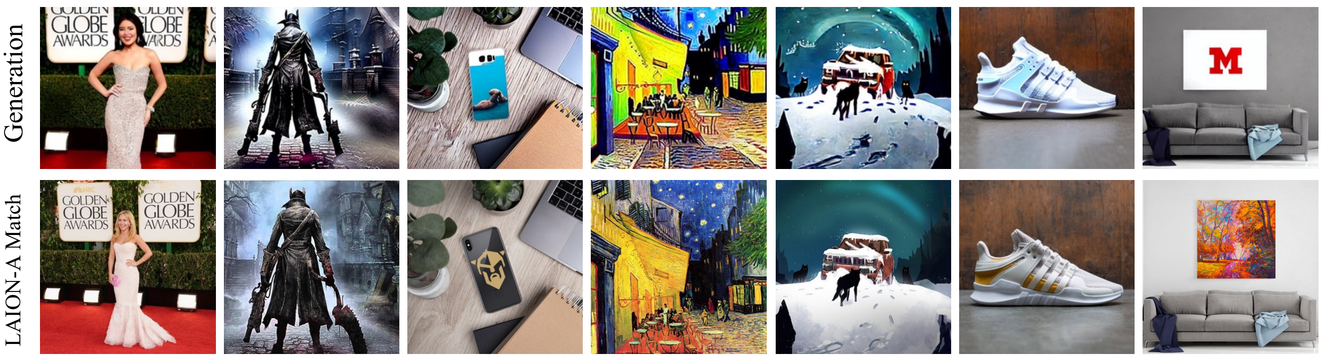


Figure 1. *Stable Diffusion* is capable of reproducing training data, creating images by piecing together foreground and background objects that it has memorized. Furthermore, the system sometimes exhibits *reconstructive* memory, in which recalled objects are semantically equivalent to their source object without being pixel-wise identical. Here, we show this behavior occurring with a range of prompts sampled from LAION, and with a hand-crafted prompt (rightmost pair). The presence of such images raises questions about the nature of data memorization and the ownership of diffusion images. Top row: generated images. Bottom row: closest matches in the LAION-Aesthetics v2 6+ set. Sometimes source and match prompts are quite similar, and sometimes they are quite different. See Fig. 6 for more examples with prompts, or the Appendix for prompts from this figure.

Abstract

Cutting-edge diffusion models produce images with high quality and customizability, enabling them to be used for commercial art and graphic design purposes. But do diffusion models create unique works of art, or are they replicating content directly from their training sets? In this work, we study image retrieval frameworks that enable us to compare generated images with training samples and detect when content has been replicated. Applying our frameworks to diffusion models trained on multiple datasets including Oxford flowers, Celeb-A, ImageNet, and LAION, we discuss how factors such as training set size impact rates of content replication. We also identify cases where diffusion models, including the popular Stable Diffusion model, blatantly copy from their training data. Project page: https://somepago.github.io/diffrep.html

1. Introduction

The rapid rise of diffusion models has led to new generative tools with the potential to be used for commercial art and graphic design. The power of the diffusion paradigm stems in large part from its reliance on simple denoising networks that maintain their stability when trained on huge web-scale datasets containing billions of image-caption pairs. These mega-datasets have the power to forge commercial models like *DALL·E* [54] and *Stable Diffusion* [56], but also bring with them a number of legal and ethical risks [7]. Because these datasets are too large for careful human curation, the origins and intellectual property rights of the data sources are largely unknown. This fact, combined with the ability of large models to memorize their training data [9, 10, 22], raises questions about the originality of diffusion outputs. There is a risk that diffusion models might, without notice, reproduce data from the training set directly, or present a collage of multiple training images.

We informally refer to the reproduction of training images, either in part or in whole, as *content replication*. In principle, replicating partial or complete information from the training data has implications for the ethical and legal use of diffusion models in terms of attributions to artists and photographers. Replicants are either a benefit or a hazard; there may be situations where content replication is acceptable, desirable, or fair use, and others where it is "stealing." While these ethical boundaries are unclear at this time, we focus on the scientific question of *whether replication actually happens with modern state-of-the-art diffusion models, and to what degree*.

Our contributions are as follows. We begin with a study of how to detect content replication, and we consider a range of image similarity metrics developed in the self-supervised learning and image retrieval communities. We benchmark the performance of different image feature extractors using real and purpose-built synthetic datasets and show that state-of-the-art instance retrieval models work well for this task. Armed with new and existing tools, we search for data replication behavior in a range of diffusion models with different dataset properties. We show that for small and medium dataset sizes, replication happens frequently, while for a model trained on the large and diverse ImageNet dataset, replication seems undetectable.

This latter finding may lead one to believe that replication is not a problem for large-scale models. However, the even larger *Stable Diffusion* model exhibits clear replication in various forms (Fig 1). Furthermore, we believe that the rate of content replication we identify in *Stable Diffusion* likely underestimates the true rate because the model is trained on a 2B image split of LAION, but we only search for matches in the smaller 12M "Aesthetics v2 6+" subset.

The level of image similarity required for something to count as "replication" is subjective and may depend on both the amount of diversity within the image's class as well as the observer. Some replication behaviors we uncover are unambiguous, while in other instances they fall into a gray area. Rather than choosing an arbitrary definition, we focus on presenting quantitative and qualitative results to the reader, leaving each person to draw their own conclusions based on their role and stake in the process of generative AI.

2. Background

Image retrieval and copy detection. The process of searching a database for images containing reference features from a source image is known as *image retrieval*. The related task of *inexact copy detection* requires high semantic similarity between the source and match [17]. Image retrieval works with image descriptors based on all types of neural networks [3, 55]. High-performance descriptors can be fine-tuned specifically for retrieval after unsupervised training [51, 52] using structure-from-motion (SfM) or contrastive objectives [14, 28]. A natural basis for image retrieval methods are self-supervised models that inherently learn strong feature descriptors, matching similar images to similar representations [11,13,15,29,31]. A particularly relevant SSL method for our purposes is DINO [12], which is shown to perform competitively on instance retrieval tasks.

Recent approaches adopt strong vision transformers as architectural backbones for retrieval [6, 19, 27, 35, 61]. Historical progress in this field is tracked by public image similarity challenges [18]. A recent SOTA approach is SSCD [49], which builds on previous work in self-supervised representation learning and optimizes a descriptor for copy detection using entropic regularization and an array of task-specific data augmentations.

In contrast to content-based retrieval techniques discussed above, there are a few style-based image retrieval methods [25, 40, 59] though it is not as popular a task as content-based retrieval.

Memorization in deep learning. While it is widely known and discussed that large models can memorize their data, there is no universally accepted definition of memorization. To ML theorists, memorization is synonymous with overfitting [2, 21, 23]. In the field of membership inference attacks, one seeks to determine whether a chosen image was part of the training set [8, 33, 66, 67]. Indeed, it has been shown that models retain a memory of the contents of their training set, particularly when training samples are repeated [67]. Note that membership inference can be done by reconstructing original training data from the model [66], although this is not the goal of most membership inference methods. The problem of explicitly reconstructing images from the training set of a classifier is known as *model inversion*, and recent research has been able to do this with both convolutional and transformer models [26, 70]. However, it is crucial to note the relationship of memorization, membership inference, inversion and *replication*: A generative model that memorizes data might allow for model inversion or only membership inference, yet the same model might never spontaneously *generate* the training data by accident.

Memorization in language. It is well known that generative language models risk replication from their training set [9, 10] and the amount of replicated data is broadly proportional to the size of the model, amount of duplication of the data point in the training set, and the amount of prompting. Interestingly, such replication behavior occurs even for models that are not overfitting to their training data [34, 63].

Diffusion models. Diffusion is a process for converting samples from a Gaussian noise distribution into samples from an arbitrary and more complex distribution, such as the distribution of natural images.

We consider several variants of diffusion models. *Stable Diffusion* is a state-of-the-art text-conditional latent diffusion model [56], trained on the LAION database [60]. The version we analyze in this work (v1.4) was initially trained on over 2B images and then fine-tuned with 600M images from the LAION Aesthetics v2 5+ subset, which is filtered for image quality. We search for matches only in the much smaller 12M LAION Aesthetics v2 6+ split to keep storage costs manageable.

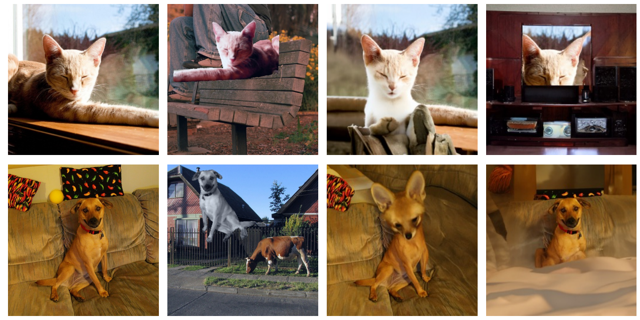
Related work. Replication behavior in GANs has been studied in a number of works. Meehan et al [41] describe a hypothesis test that discerns whether generated images are on average closer to the training data than a random sample from a hold-out set. Note that this test is at the population level, and is not designed to flag individual instances of replication. Feng et al. [24] study the conditions that lead GANs to replicate training data. They look for copies in pixel-space and find that such replications are inversely proportional to dataset complexity and dataset size. Webster et al [66] show on face datasets that GANs can occasionally replicate. Interestingly, these models can produce novel images of known identities from the training data without making verbatim copies. FID scores for ranking GANs favor models that memorize training data [4], leading toward a search for measures of generalization without memorization [30]. This includes "authenticity scores" that detect replication [1], but only in the form of noisy pixel-by-pixel copies of the training data. Similarly, authors of large-scale diffusion models have investigated image replication themselves [42], reducing replication through training data de-duplication, and checking for simple nearest-neighbor matches.

3. What Counts as Replication?

There are many different notions of *replication* from creative work, but we will narrow our scope for the purpose of designing a detection system for replicated content. We consider the following (informal) definition:

We say that a generated image has replicated content if it contains an object (either in the foreground or background) that appears identically in a training image, neglecting minor variations in appearance that could result from data augmentation.

We focus on object-level similarity because it is likely to be the subject of intellectual property disputes. We also discount minor differences in appearance that can be explained by data augmentation as these variations would typically not be relevant to a copyright claim. An alternative notion is style-wise or semantic similarity. We do not focus on such definitions here as they are highly subjective, typically are not considered an infringement of intellectual property, and also because many images lack a well-defined style (e.g., natural, unfiltered images from a standard camera).



(a) (b) (c) (d)

Figure 2. Synthetic datasets. (a) Original images. (b) Segmix generation. (c) Diagonal outpainting. (d) Patch outpainting. Please refer to Section 4 for more details.

4. Detecting Content Replication

Our goal is to construct a system to detect replication as defined above. To find a powerful system, we consider 10 different prototypes of feature extractors drawn from the SSL and image retrieval literature. We compare and contrast these methods using 10 different datasets that we curate for measuring the performance of replication detectors.

Synthetic datasets. There are currently no existing labeled datasets that capture our notion of replication as defined above. Thus, we create 5 synthetic datasets. Our *IN-Cutmix* dataset is built by pasting random square patches from one image into a random location in another. The size of the pasted patch is randomly chosen. We use ImageNet as the source for base images [58]. Our *IN-Dif-Patch* dataset is created by masking 80% of the image except for a random square patch and then outpainting the rest of the pixels using the method proposed by Lugmayr et al [39].

In the above two datasets, the replicated content lies inside a square patch. Vision transformer models naturally rely on square patches, and so we create the *IN-Dif-Diagonal* dataset by masking a random triangular half of an ImageNet image (above or below the diagonal) and using diffusion to outpaint the masked region, resulting in an image that shares half the content of the original.

Since real world objects may have irregular shapes, we next use the segmentation masks from the MS COCO [36] and Pasal VOC [20] datasets to generate the synthetic data. For a random query image, we choose either a single object or its background. We then apply a plethora of augmentations (flips, blur, autocontrast, solarize, colorjitter) to the selected region before pasting it into another random image. If a foreground object is chosen, we also resize and reposition the object at random. We call these challenging datasets *MSCOCO Segmix* and *VOC Segmix*. See Fig. 2.

Real datasets. Similarity-based image retrieval is closely related to copy detection, although the matching criteria is less stringent for retrieval. We choose 5 im-

Table 1. We present the mAP scores for all 10 models across 10 datasets. The first five datasets are real and the next five are synthetic. In the last column we show the average rank of each model across datasets. We categorized the models based on the style of training. The categories are as follows: CD/IR - Copy Detection/ Instance Retrieval, PT - Pre-Trained, SSL - Self-Supervised Learning. Refer Section 4 for more details on models, datasets and the metric. mAP higher the better. Average rank lower the better.

Type	Method	rOxford5k ↑	rParis6k ↑	CUB-200 ↑	GPR1200 ↑	INSTRE ↑	MSCOCO Segmix ↑	VOC-Segmix ↑	IN-Cutmix ↑	IN-Dif-Diagonal ↑	IN-Dif-Outpaint ↑	Average Rank ↓
CD/IR	Multigrain [6], ResNet-50	22.87	44.74	3.67	37.09	56.72	27.77	25.8	67.54	79.44	24.91	5.9
	SSCD [49] , ResNet-50	30.16	45.75	2.00	31.42	53.54	67.04	65.82	89.78	99.91	96.11	**4.2**
PT	ViT [16] S/16, IN1k	31.24	61.5	13.12	40.44	54.11	23.21	22.42	61.99	48.36	14.25	5.5
	ViT-B/16, IN12k	13.14	30.43	4.24	16.63	29.15	18.15	15.61	52.74	49.69	10.18	9.2
	ViT-B/16, CLIP [53] on LAION [60]	39.92	68.92	8.6	62.13	73.19	20.37	17.91	59.5	47.54	8.76	5.4
	Swin Transformer [37], Base, IN1k	40.06	72.07	15.49	54.09	68.46	24.51	24.31	74.79	40.74	14.75	**4.1**
SSL	MoCo [15], ViT-B/16	30.25	51.6	4.94	37.98	51.88	36.41	32.9	65.98	59.12	20.61	5.1
	MoCo, ViT-B/16 + CutMix [71]	25.01	46.73	3.44	32.23	48.58	32.83	26.11	55.74	62.88	46.96	6.5
	VicRegL [5], ResNet-50	28.4	53.79	3.02	34.95	50.98	40.58	37.76	69.74	80.02	40.93	5.0
	DINO [12], ViT-B/16, split-product	32.14	45.43	5.76	29.41	50.06	46.42	45.29	93.53	98.92	95.86	**4.1**

age retrieval datasets with high diversity. *Oxford* [47] and *Paris* [48] are geographic landmark datasets where query and gallery images contain the same building. We use the cleaned-up version with corrected labels [50]. *INSTRE* [65] contains objects like toys or irregularly-shaped products placed in different locations and conditions. *GPR1200* is a general-purpose content retrieval dataset with 1200 classes sampled from other datasets such as Google Landmarks V2 [68], Stanford Online Products [44], IMDB-WIKI [57] and others. Caltech-UCSD Birds-200 or *CUB-200* [64] is a dataset with fine-grained classes of birds in different backgrounds, poses and lighting conditions.

Models. Several recent self-supervised (SSL) methods are competitive with supervised retrieval techniques. To the best of our knowledge, no rigorous study exists that compares multiple SSL models to retrieval specialist models across multiple datasets. Our study uses the following candidate models and training methodologies.

MultiGrain [6] trains a retrieval model with both classification and retrieval triplet loss. We used the best-performing ImageNet pre-trained ResNet-50 checkpoint from the official repo. SSCD [49] is a self-supervised copy detection method trained in the style of SimCLR [13] using InfoNCE loss [45], entropy regularization on latent space representations, and many strong augmentations. We used the official ResNet-50 checkpoint trained on ImageNet. Some established methods use pre-trained models as backbones to perform image retrieval tasks [14]. Hence we also evaluate 4 models from timm [69] that are trained in a supervised fashion. ViT-Small/16 [16] pre-trained on ImageNet, ViT-Base/16 pre-trained on ImageNet-21k, ViT-Base/32 image encoder from the CLIP [53] model trained on LAION [60], and finally a Swin-Transformer [37] with base patch 4 and window 7 trained on ImageNet.

Lastly, we explored 3 self-supervised models. First is a ViT-Base/16 variant trained with the DINO [12] framework. We also consider ResNet-50 from VICRegL [5]. Finally we consider ViT-Base/16 from MoCo v3 [15], and a variant of MoCo v3 that we fine-tune for 50 epochs with CutMix [71] as an additional augmentation with the goal of boosting its copy detection performance.

Computing the similarity. It is common to compare two images via the inner product of feature vectors (either the `[CLS]` token or average-pooled representations) [12, 14, 15]. Inner product metrics measure global, rather than local similarity. This is because inner product spaces are metric spaces and thus satisfy the triangle inequality. To see why this is a problem, consider an example in which generated image $\mathcal{I}_{\text{gen}}$ contains a car and a tree directly stolen from two unrelated images $\mathcal{I}_{\text{car}}$ and $\mathcal{I}_{\text{tree}}$, respectively. Then we would like $d(\mathcal{I}_{\text{gen}}, \mathcal{I}_{\text{car}})$ and $d(\mathcal{I}_{\text{gen}}, \mathcal{I}_{\text{tree}})$ to be very small indicating replication. But by the triangle inequality, the two unrelated images satisfy $d(\mathcal{I}_{\text{car}}, \mathcal{I}_{\text{tree}}) \leq d(\mathcal{I}_{\text{gen}}, \mathcal{I}_{\text{car}}) + d(\mathcal{I}_{\text{gen}}, \mathcal{I}_{\text{tree}})$, and are also scored as similar even though they share nothing.

To bypass this potential problem, we implement a *split-product* metric that breaks each feature vector into chunks, computes inner products between corresponding chunks, and returns the maximum across these inner products. In vision transformers, we use the representation corresponding to each token as a chunk since they are more local in nature than the `[CLS]` token. Under this strategy, if $d(\mathcal{I}_{\text{gen}}, \mathcal{I}_{\text{car}})$ and $d(\mathcal{I}_{\text{gen}}, \mathcal{I}_{\text{tree}})$ are small, then for each of these two image pairs, at least one such feature vector chunk must yield a high inner product. However, the locations of these two chunks, each corresponding to one of the image pairs, may differ so that $d(\mathcal{I}_{\text{car}}, \mathcal{I}_{\text{tree}})$ may remain large. We test both the split-product and standard inner product metric and find that both can return suitable, and often differing, matches. Qualitatively, the split product metric is a more semantic measure of similarity. As expected, the inner product metrics enforces a stricter notion of pixel-wise similarity.

4.1. Choosing the Best Replication Detector

We measure model performance using mean-Average-Precision or *mAP* [46]. Tab. 1 shows mAP scores for all the models across different datasets. We also present average ranks of each model averaged across all datasets (Lower is

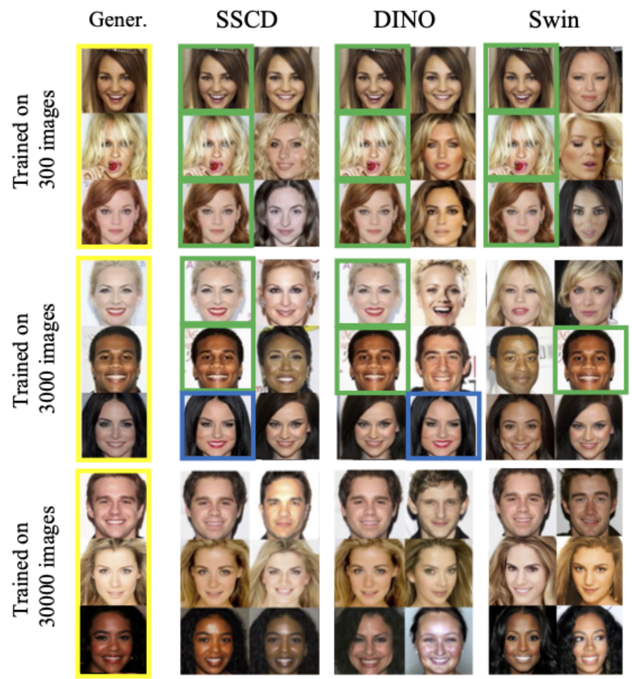


Figure 3. The top two matches (according to different feature extractors) for generations across diffusion models trained on datasets of size 300, 3000 and 30000 (whole dataset). Across the board, we can see full replication in the first 2 models (indicated in green). Very close but not exact copies are indicated in blue. However in the model trained on the whole dataset, the first matches are *very similar* but not the same. Refer to Section 5 for more details.

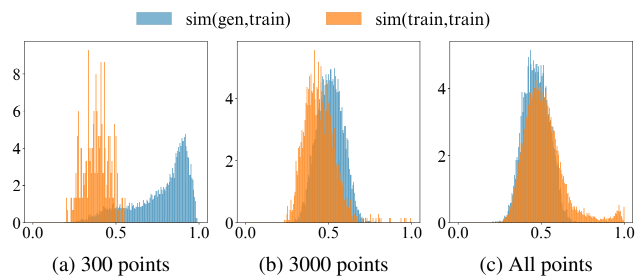


Figure 4. The histograms of top-1 similarity scores between generations and the training data and the top-1 self-similarity scores of training data across 3 diffusion models trained with different amounts of training data. Varying amount of training data is used in each plot but 10000 generations are used across all.

better). DINO [12] with split-product performed the best on average across all 10 datasets. For real datasets, the winner is Swin Transformer [37], and for synthetic datasets SSCD [49] does best. For the rest of the paper, we focus our studies on SSCD, Swin, and DINO (with split-product) as best performing methods in Tab. 1.

5. Do Diffusion Models Copy?

In this section, we methodically explore diffusion models trained on different datasets with varying amounts of training data. We observe that the diffusion models trained on smaller datasets tend to generate images that are copied from the training data. The amount of replication reduces as we increase the size of the training set.

Experimental setup. We train Denoising Diffusion Probabilistic Models (DDPM) [32] with a discrete denoising scheduler on various datasets using the HuggingFace implementation. For Celeb-A [38], we train two models on 300 and 3000 training images. We also use the full dataset pre-trained checkpoint from the official repository. For Oxford Flowers [43], we train models on 100, 1083 (top 5 classes), and 8189 (complete dataset) images. We train all models with random horizontal flip and random crop augmentations. In all cases, we train the models until generations appear to be high quality, for at least 300k steps and until the FID [62] scores are lower than 50. We do quantitative analysis on 10000 generations for Celeb-A and 5000 generations for Oxford Flowers.

Finding matches. For each generated image, we search the training set using dot products between its features and training samples (except for DINO which uses split-product). All the generations used in figures are from the 20 generated images with highest top-1 similarity scores for standard diffusion models, and are among images with similarity > 0.5 for *Stable Diffusion*.

Qualitative observations. Fig. 3 and Fig. 13 show generated images and their corresponding top matches from the training dataset. We consider diffusion models (DDPM) trained with varying amounts of training data. In the case of Celeb-A, diffusion models trained on 300 and 3000 images blatantly copy from their training images. However, when the model is trained on the whole dataset, generations may appear that are similar to training samples, but not identical. We observe similar trends in diffusion models trained on the Oxford Flowers dataset as well (Appendix Fig. 13).

Quantitative observations. To further complement our visual inspection, we can also examine the distribution of similarity scores between generated images and training samples. Fig. 4 contains histograms of similarity scores between generations and their best match from the training data. As a baseline, we also draw random training images and compute the similarity with their closest match from the remaining training images. If most scores between generated and training images lie to the right of this baseline, then the model is generating images that are closer to their training samples than the training samples are to each other.

Most samples generated by the 300-sample model are extremely similar to the training data, having very high similarity scores. However, the histogram's mass shifts drastically to left when we train the model instead on 3000 points. We do see blatant copies from this model too, but this phenomenon occurs infrequently. The histograms of similarity scores computed using the full dataset model are highly overlapping. This strong alignment indicates that the model is not, on average, copying its training images any more than its training images are copies of each other. The histogram of generated images (blue) no longer has a long right tail, indicating that the model is unlikely to generate exact copies of its training samples. Note that a small proportion of the dataset self-similarity scores in Fig. 4 (c) are greater than 0.9, indicating that there are repetitions or near repetitions in the training data.

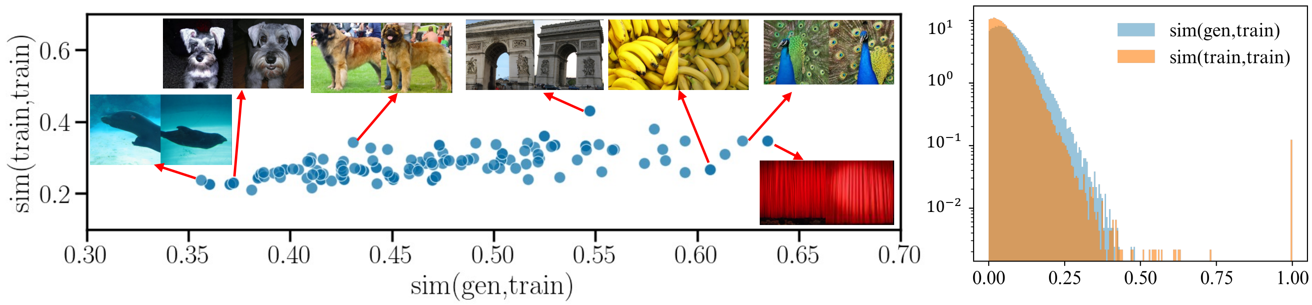


Figure 5. (Left) For each individual class, we plot top-1 similarity between a generation and training data vs the mean of the top-1 similarity scores of the training data to itself. We also show a few images for some interesting classes. Left image is generation and right image is closest match. (Right) We show histograms of top-1 similarity scores for all the generations in 100 classes and the self-similarity scores of training data (within the class).

6. Case Study: ImageNet LDM

Experimental setup. In the previous section, we observed copying behavior when diffusion models are trained on small datasets, and the rate of copying decreases as models are trained on more data. In this section, we extend our study to an off-the-shelf class conditional Latent Diffusion Model [56] trained on ImageNet. We search for copying both at the class and the population level. We use the pre-trained model from the official repo.[1] We randomly choose 100 classes and generate 1000 samples per class (comparable to the size of training data per class in ImageNet).

Observations. Qualitatively, we observe no significant copying in any of the generations by this model. In Fig. 5 (a), we present a scatter plot with x-axis showing the maximum similarity scores observed between generations of a given class and ImageNet training samples. On the y-axis, we show the average similarity scores per class observed between training samples in that class. For a few interesting points, we also show the corresponding generation and the top match in the training data. We see the similarity scores never cross 0.65, and when we manually sift through the high similarity score examples in each of the 100 classes, they are **very similar** but never exact copies, and may be explained by low intra-class diversity

We also check if there is a relationship between the intra-class diversity and similarity scores, and indeed classes with higher self-similarity scores on average have higher maximum similarity score amongst matches with generated samples. Specifically, the points in the scatter plot have a correlation of 0.6 and the line of best fit has slope 0.39. The classes with the highest similarity between generated images and training data are `theater curtain`, `peacock`, and `bananas`. Meanwhile `sea lion`, `bee`, and `swing` are at the lower end of the spectrum. In Fig. 5 (b), we consolidate our results across the classes into a histogram of similarity scores between generations and matches from the training set and the similarity scores of training images with matches from the remaining training samples. The average similarity scores are relatively low for this dataset as well as for this diffusion model showing that the chance of replication is very low.

7. Case Study: *Stable Diffusion*

In this section, we evaluate *Stable Diffusion* v.1.4 [56], which was trained on the publicly available LAION [60] dataset. Since it is computationally expensive to store and search 2 billion+ images, we narrow our search scope to the smaller LAION Aesthetics v2 6+ dataset which has 12M images and is a **subset** of images that were used for the final rounds of training. We load the model and the checkpoints via HuggingFace [2].

In the first experiment, we randomly sample 9000 images, which we call *source images*, from LAION Aesthetics 12M and retrieve the corresponding captions. Then, we generate synthetic images by passing these captions into *Stable Diffusion*. We study the top-1 matches, which we call *match images*, for each generated sample.

We attempt to answer the following questions in this analysis. 1) Is there copying in the generations? 2) If yes,

[1] `github.com/CompVis/stable-diffusion`

[2] `huggingface.co/CompVis/stable-diffusion-v1-4`

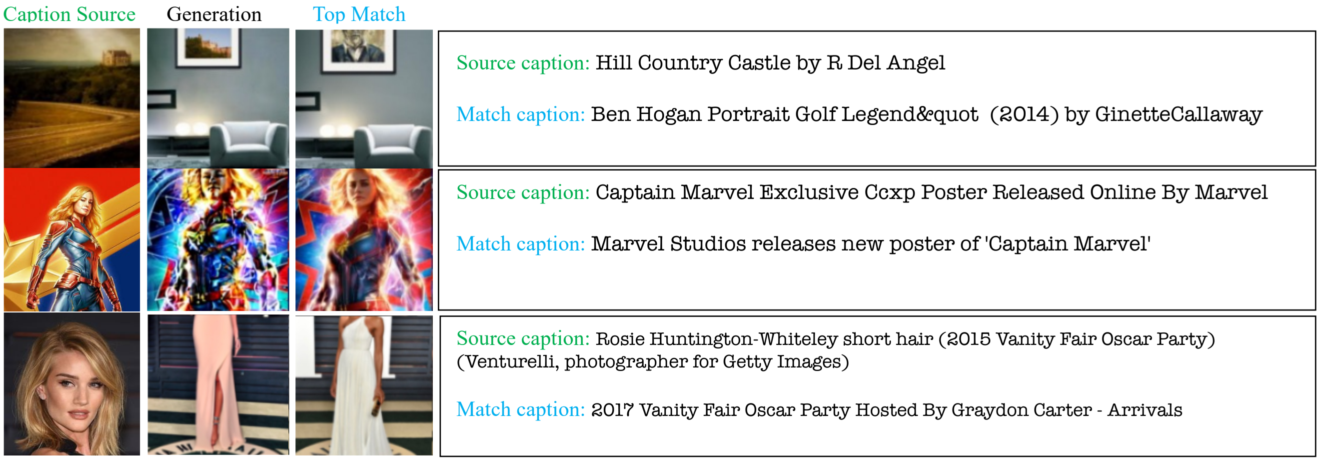


Figure 6. Selected *Stable Diffusion* generations using captions sampled from LAION images, with similarity score ≥ 0.5. Also see Fig. 12

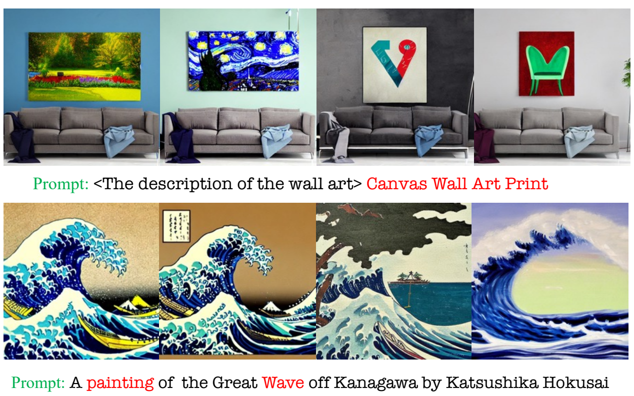


Figure 7. Including the phrase highlighted in red into a random prompt for *Stable Diffusion* leads to exact replications of the sofa (top row) and wave shape (bottom row).

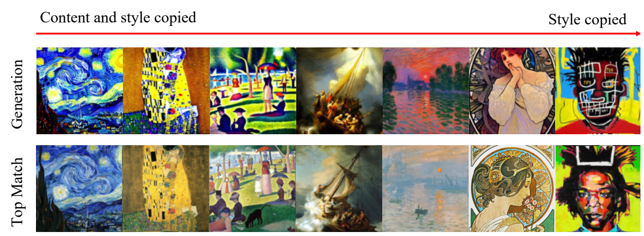


Figure 8. *Stable Diffusion* replicates pixel-level details, structures, and styles of well known paintings.

what kind of copying? 3) Does a caption sampled from the training set produce an image that matches its original source? 4) Is content replication behavior associated with training images that have many replications in the dataset?

In previous experiments, we observed that DINO with split-product is slightly better than SSCD at finding copies. But we use SSCD to study *Stable Diffusion* because of its much faster speed when crawling through the large 12M image dataset. We constructed visualizations in this section by choosing from images with an SSCD similarity > 0.5.

Observations. In Fig. 6, we visualize a few instances of copying found in samples generated by *Stable Diffusion*. We choose them from a small set of points ($\approx$ 170 images) whose top-1 similarity scores are > 0.5 (top 1.88 percentile). Above this 0.5 threshold, we observe a significant amount of copying. The first row (where only the painting changed) shows verbatim usage of an object and background. The 3rd row show local copying where only the background is recycled from the training set. We see similar trends in other images with high similarity scores. We refer the reader to Appendix Fig. 12 for more examples.

While all synthetic images were generated using captions sourced from LAION, none of the generations match their respective source image. In fact, sometimes the caption of the source image is not representative of the source image content, and the generation is quite different from the source. This behavior can be seen in the first row of Fig. 6.

In those 170 images, we find instances where replication behavior is highly dependent on key phrases in the caption. We show two examples in Fig. 7 and highlight the key phrase in red. For the first row, the presence of the text `Canvas Wall Art Print` frequently ($\approx$ 20% of the time) results in generations containing a particular sofa from LAION (also see Fig 1). Similarly, the second row shows various generations by tweaking the prompt `A painting of the Great Wave off Kanagawa by Katsushika Hokusai`. We gradually remove words until only `painting` and `wave` remain. All of the generations have a wave structure that resembles the original painting. We also notice instances of generations where style is copied rather than content. This can be explicitly seen when the name of an artist is used in the

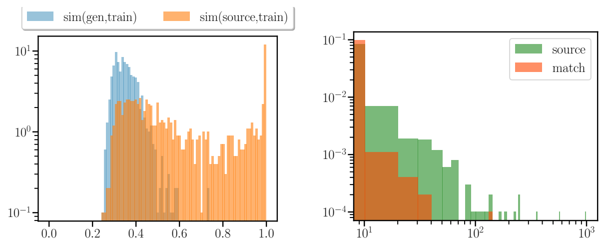


Figure 9. *Stable Diffusion* density plots. (Left) Similarity scores of 1000 random generations with train data and the corresponding 1000 caption source images with train data. (Right) Histogram of number of training set duplications for 1000 random source images and 1000 match images. Random data (source) is relatively repeated more than matched images.

Figure 10. *Stable Diffusion* generates the painting "The Scream." This image is within the 600M image LAION-Aesthetics-5+ split that was used for training, but is not within the 12M image LAION-Aesthetics-6+ split that we searched in this study.

generation prompt. We generate many paintings with the prompt style "`<Name of the painting>` by `<Name of the artist>`". We tried around 20 classical and contemporary artists, and we observe that the generations frequently reproduce known paintings with varying degrees of accuracy. In Fig. 8, as we go from left to right, we see that content copying is reduced, however, style copying is still prevalent. We refer the reader to the appendix for the exact prompts used to generate Figs. 7 and 8.

Role of caption sampling. Several of our studies sample random captions from LAION itself, and this may lead to a replica of the caption source image. Fig. 9 (left) shows histograms of top-1 similarity scores for 1000 random generations, and the top-1 similarity scores of the 1000 corresponding images from which we sourced the captions. We find the mean generation similarity scores to be much lower than those of the source images, indicating that this method of caption sampling does not typically produce a replica of the caption source.

Still, captions from LAION will sometimes contain "key phrases" that are closely associated with dataset images and therefore conjure a memorized image. For this reason, LAION-based caption sampling may lead to higher rates of data replication than other sampling methods. It is difficult to correct this bias in a scientific way, as there is no baseline sampler for "typical" captions. Furthermore, captions used by experienced diffusion users will often exploit powerful key phrases (e.g. "art station," "35mm," or the name of an artist) that widely appear in LAION.

Role of duplicate training data. Many LAION training images appear in the dataset multiple times. It is natural to suspect that duplicated images have a higher probability of being reproduced by *Stable Diffusion*. Fig. 9 (right) shows a histogram of how many times a training image is duplicated in LAION-Aesthetics, where "duplicate" is defined as having SSCD score > 0.95. We plot a histogram for two populations. First, the 1000 sampled source images. Second, we generate 1000 synthetic images, search for their closest match in the training set, and plot the duplication histogram for these "match" images. Surprisingly, a typical random image from the dataset is duplicated 11.6 times, which is *more* often than a typical matched image, which is duplicated 3.1 times. However, if we look only at very close matches ($> .5$ SSCD), these match images are replicated on average 34.1 times – far more often than a typical image. It seems that replicated content tends to be from training images that are duplicated more than a typical image.

8. Limitations & Conclusion

The goal of this study was to evaluate whether diffusion models are capable of reproducing high-fidelity content from their training data, and we find that they are. While most of the generations from large-scale models do not contain copied content, a non-trivial amount of copying does occur; *Stable Diffusion* images with dataset similarity $\geq .5$, as depicted in Figs. 1 and 6, account for approximate 1.88% of our random generations.

Note, however, that our search in *Stable Diffusion* only covered the 12M images in the LAION Aesthetics v2 6+ dataset. The model was first trained on ~ 2 billion images, and the dataset we searched in our study is a small subset of this fine-tuning data, comprising less than 0.6% of the total training data. Examples certainly exist of content replication from sources outside the 12M LAION Aesthetics v2 6+ split – see Fig 10. Furthermore, replication very likely exists that our retrieval method is unable to identify. For both of these reasons, the results here systematically underestimate the amount of replication in *Stable Diffusion* and other models. Lastly, we refer the reader to Appendix A for a discussion on potential causes of replication.

9. Acknowledgements

This work was supported by the ONR MURI program, Office of Naval Research (N000142112557), the AFOSR MURI program, DARPA GARD (HR00112020007), and the National Science Foundation (IIS-2212182 & DMS-1912866). Further support was provided by Capital One Bank.

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 290–306. PMLR, June 2022. 3

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 2

[3] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *European conference on computer vision*, pages 584–599. Springer, 2014. 2

[4] Ching-Yuan Bai, Hsuan-Tien Lin, Colin Raffel, and Wendy Chi-wen Kan. On Training Sample Memorization: Lessons from Benchmarking Generative Modeling with a Large-scale Competition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 2534–2542, New York, NY, USA, Aug. 2021. Association for Computing Machinery. 3

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *arXiv preprint arXiv:2210.01571*, 2022. 4

[6] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. Multigrain: a unified image embedding for classes and instances. *arXiv preprint arXiv:1902.05509*, 2019. 2, 4

[7] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arxiv:2110.01963[cs]*, Oct. 2021. 1

[8] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, May 2022. 2

[9] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying Memorization Across Neural Language Models. *arxiv:2202.07646[cs]*, Feb. 2022. 1, 2

[10] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021. 1, 2

[11] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 2

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2, 4, 5

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 4

[14] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2, 4

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 2, 4

[16] 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

[17] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 1–8, 2009. 2

[18] Matthijs Douze, Giorgos Tolias, Ed Pizzi, Zoë Papakipos, Lowik Chanussot, Filip Radenovic, Tomas Jenicek, Maxim Maximov, Laura Leal-Taixé, Ismail Elezi, et al. The 2021 image similarity dataset and challenge. *arXiv preprint arXiv:2106.09672*, 2021. 2

[19] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644*, 2021. 2

[20] Mark Everingham, SM Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 3

[21] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020. 2

[22] Vitaly Feldman and Chiyuan Zhang. What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation. *arxiv:2008.03703[cs, stat]*, Aug. 2020. 1

[23] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 2

[24] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix M Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021. 3

[25] Elena Garces, Aseem Agarwala, Diego Gutierrez, and Aaron Hertzmann. A similarity measure for illustration style. *ACM Transactions on Graphics (TOG)*, 33(4):1–9, 2014. 2

[26] Amin Ghiasi, Hamid Kazemi, Steven Reich, Chen Zhu, Micah Goldblum, and Tom Goldstein. Plug-In Inversion: Model-Agnostic Inversion for Vision with Data Augmentations. In *Proceedings of the 39th International Conference*

on Machine Learning, pages 7484–7512. PMLR, June 2022. 2

[27] Yunchao Gong, Liwei Wang, Ruiqi Guo, and Svetlana Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European conference on computer vision*, pages 392–407. Springer, 2014. 2

[28] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 2

[29] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 2

[30] Ishaan Gulrajani, Colin Raffel, and Luke Metz. Towards GAN Benchmarks Which Require Generalization. In *International Conference on Learning Representations*, Feb. 2022. 3

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 2

[32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 5

[33] Hailong Hu and Jun Pang. Membership Inference Attacks against GANs by Leveraging Over-representation Regions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pages 2387–2389, New York, NY, USA, Nov. 2021. Association for Computing Machinery. 2

[34] Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. Measuring Forgetting of Memorized Training Examples. *arxiv:2207.00099[cs]*, June 2022. 2

[35] Young Kyun Jang and Nam Ik Cho. Self-supervised product quantization for deep unsupervised image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12085–12094, 2021. 2

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[37] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 4, 5

[38] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3

[40] Shin Matsuo and Keiji Yanai. Cnn-based style vector for style image retrieval. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 309–312, 2016. 2

[41] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020. 3

[42] Alex Nichol, Aditya Ramesh, Pamela Mishkin, Prafulla Dariwal, Joanne Jang, and Mark Chen. DALL·E 2 Pre-Training Mitigations, June 2022. 3

[43] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008. 5

[44] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 4

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[46] Florent Perronnin, Yan Liu, and Jean-Michel Renders. A family of contextual measures of similarity between distributions with application to image retrieval. In *2009 IEEE Conference on computer vision and pattern recognition*, pages 2358–2365. IEEE, 2009. 4

[47] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 4

[48] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 4

[49] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 2, 4, 5

[50] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5706–5715, 2018. 4

[51] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016. 2

[52] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018. 2

[53] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1

[55] Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4(3):251–258, 2016. 2

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 6

[57] Rasmus Rothe, Radu Timofte, and LV Gool. Imdb-wiki-500k+ face images with age and gender labels. *Online] URL: https://data. vision. ee. ethz. ch/cvl/rrothe/imdb-wiki*. 4

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 3

[59] Dan Ruta, Saeid Motiian, Baldo Faieta, Zhe Lin, Hailin Jin, Alex Filipkowski, Andrew Gilbert, and John Collomosse. Aladin: all layer adaptive instance normalization for fine-grained style similarity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11926–11935, 2021. 2

[60] 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, 4, 6

[61] Chull Hwan Song, Jooyoung Yoon, Shunghyun Choi, and Yannis Avrithis. Boosting vision transformers for image retrieval. *arXiv preprint arXiv:2210.11909*, 2022. 2

[62] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. 2015. *arXiv preprint arXiv:1512.00567*, 2015. 5

[63] Kushal Tirumala, Aram H Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *arXiv preprint arXiv:2205.10770*, 2022. 2

[64] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 4

[65] Shuang Wang and Shuqiang Jiang. Instre: a new benchmark for instance-level object retrieval and recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(3):1–21, 2015. 4

[66] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces. *arxiv:2107.06018[cs]*, July 2021. 2, 3

[67] Yuxin Wen, Arpit Bansal, Hamid Kazemi, Eitan Borgnia, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Canary in a Coalmine: Better Membership Inference with Ensembled Adversarial Queries. *arxiv:2210.10750[cs]*, Oct. 2022. 2

[68] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020. 4

[69] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 4

[70] Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. Dreaming to Distill: Data-free Knowledge Transfer via DeepInversion. *arxiv:1912.08795[cs, stat]*, June 2020. 2

[71] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 4