Use of Copyrighted Material in the Training of Stable Diffusion

Michael Frank

Introduction

This paper is written in response to the United States Copyright Office's request for comments on Generative AI. Specifically, it seeks to address the following questions posed by the Office:

- *6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?*

- *6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?*

These questions will be answered in the context of StabilityAI and RunwayML's development of image generation models Stable Diffusion 1.x and StabilityAI's development of Stable Diffusion 2. (*Stability AI no longer discloses the source of its training data for their latest model, SDXL.*) These models indicate their use of the LAION5B dataset, a listing of 5 billion text and image url pairs, as evidenced by their documentation.

The curators of the LAION datasets and developers of Stable Diffusion demonstrate a bias towards works of "aesthetic quality" - not all 5 billion images are weighted equally (or used at all) in the training of Stable Diffusion models.

Beyond the raw information contained inside of LAION5B, I have also quantified the amount of image urls according to the hosts from which they originate. Though the website information may not indicate a possession of copyright or valid licensing, it may provide some insight into the intended use of the imagery. These last data points are rather large and will be attached as separate documents in my submission form.

For the ease of browsing this data, I've created a website -- https://whatsinlaion.com/ that contains these listings as well as a gallery view of the image urls contained inside of LAION's improved aesthetics dataset.

Collection and Curation

According to the Stable Diffusion 1.x documentation, Stable Diffusion was trained on "laion2b-en", "laion-high-resolution", "laion-improved-aesthetics" and "laion-aesthetics v2 5+". These are references to datasets created by the organization LAION, "a non-profit organization with members from all over the world, aiming to make large-scale machine learning models, datasets and related code available to the general public." (Laion, n.d.)

Entries in the LAION datasets contain the following data points:

1. URL - the `src=` web address of the image
2. TEXT - the 'alt=' text used to describe the image
3. WIDTH - width of the image in pixels
4. HEIGHT - height of the image in pixels
5. similarity - calculated similarity between TEXT and image content as calculated by CLIP
6. punsafe - the probability that the image is NSFW

Additionally, the 'aesthetic' datasets add another data point:

7. AESTHETIC_SCORE - a 1 to 10 ranking provided by a model trained to rank images for visual quality.

These datasets are publicly available to view and download from developer communities such as HuggingFace. They are comprised of the following datasets:

https://huggingface.co/datasets/laion/laion2B-en
https://huggingface.co/datasets/laion/laion2B-multi
https://huggingface.co/datasets/laion/laion1B-nolang
https://huggingface.co/datasets/laion/laion-high-resolution
https://huggingface.co/datasets/ChristophSchuhmann/improved_aesthetics_4.5plus

Common Crawl

"To create image-text pairs, we parse through WAT files from Common Crawl and parse out all HTML IMG tags containing an alt-text attribute." (Beaumont 2022)

LAION collected this information using web archive data from the project Common Crawl,a 501(c)(3) non–profit that makes "wholesale extraction, transformation and analysis of open web data accessible to researchers." (CC, n.d.)

Common Crawl is a web crawler, or, an “internet bot that systematically browses the World Wide Web and that is typically operated by search engines for the purpose of Web indexing.” (“Web crawler”, n.d.) Common Crawl data is stored in a Web ARChive (WARC) Format and contains raw HTTP Responses, which among other responses, will include the full html document used to display a website. An example of this data can be seen here:

https://gist.github.com/Smerity/e750f0ef0ab9aa366558#file-bbc-warc

In particular the kind of information used to compile the LAION datasets would be found in the ‘<img>’ tags of the .html. As seen here on line 717 of the same example:

https://gist.github.com/Smerity/e750f0ef0ab9aa366558#file-bbc-warc-L717

For an example of a utility that extracts these image tags, LAION co-author Romain Beaumont wrote a python library, cc2dataset, that may be used to perform this extraction. (“cc2dataset”, n.d.)

Collection of Training Data

LAION datasets are created at a massive scale and contain information from all over the internet. Their content raises questions of copyright, publicity, and privacy. Additionally, the size of the images themselves is estimated to be petabytes of data! (The size of just the LAION2B-en listings is around 321 gb.)

In an attempt to avoid problems of liability, as well as the problems of data storage, LAION only distributes the web addresses of the image, not the actual image itself. This means that machine learning developers utilizing the LAION datasets (such as StabilityAI) must download the images directly from the hosts before beginning training.

When downloading images using these indexes, copies are necessarily made (however transient) to facilitate the AI training process. These copies are used without approval of the original rights holder, or the approval of the web provider. Furthermore, the existence of a publicly accessible image on the internet does not imply that its original use is licensed or authorized.

Bias for Aesthetic Quality

"Eh when you release a deduped and flattened model you get crapped on a bit. Folk really like aesthetic tuned ones" ("@EMostaque - Mar 13, 2023" 2023)

"1.x was unfiltered LAION 2b (English), then moved through aesthetic and other.

2.x was filtered on nsfw, watermarks and other things and a much smaller dataset, as well as deduplicating etc." ("@EMostaque - Jan 19, 2023" 2023)

In common parlance, it's often stated that Stable Diffusion models are trained on LAION-5B, implying that Stability AI used the total of 5 billion images contained in the dataset without discrimination. In practice, we can examine the Stable Diffusion model cards to see how training was performed. From the Stable Diffusion 1.5 model card:

Currently six Stable Diffusion checkpoints are provided, which were trained as follows.

- `stable-diffusion-v1-1`: 237,000 steps at resolution `256x256` on laion2B-en. 194,000 steps at resolution `512x512` on laion-high-resolution (170M examples from LAION-5B with resolution `>= 1024x1024`).
- `stable-diffusion-v1-2`: Resumed from `stable-diffusion-v1-1`. 515,000 steps at resolution `512x512` on "**laion-improved-aesthetics**" (a subset of laion2B-en, filtered to images with an original size `>= 512x512`, estimated aesthetics score `> 5.0`, and an estimated watermark probability `< 0.5`. The watermark estimate is from the LAION-5B metadata, the aesthetics score is estimated using an improved aesthetics estimator).
- `stable-diffusion-v1-3`: Resumed from `stable-diffusion-v1-2` - 195,000 steps at resolution `512x512` on "**laion-improved-aesthetics**" and 10 % dropping of the text-conditioning to improve classifier-free guidance sampling.
- `stable-diffusion-v1-4` Resumed from `stable-diffusion-v1-2` - 225,000 steps at resolution `512x512` on "**laion-aesthetics v2 5+**" and 10 % dropping of the text-conditioning to improve classifier-free guidance sampling.
- `stable-diffusion-v1-5` Resumed from `stable-diffusion-v1-2` - 595,000 steps at resolution `512x512` on "**laion-aesthetics v2 5+**" and 10 % dropping of the text-conditioning to improve classifier-free guidance sampling.
- `stable-diffusion-inpainting` Resumed from `stable-diffusion-v1-5` - then 440,000 steps of inpainting training at resolution 512x512 on **"laion-aesthetics v2**

> **5+"** and 10% dropping of the text-conditioning. For inpainting, the UNet has 5 additional input channels (4 for the encoded masked-image and 1 for the mask itself) whose weights were zero-initialized after restoring the non-inpainting checkpoint. During training, we generate synthetic masks and in 25% mask everything.

From this file we can read the following:

Stable Diffusion 1.1 was trained on subsets laion2B-en and laion-high-resolution. Subsequent versions of Stable Diffusion 1.x were trained using versions of 'laion-improved-aesthetics' and 'laion-aesthetics v2 5+'.

This is echoed by LAION's aesthetics page, which states, "We provided the dataset to the CompViz team led by Robin Rombach and Patrick Esser. They used the 5+ subset to train Stable Diffusion V1 model." (Schuhmann 2022)

The Stable Diffusion 2.1 model card also demonstrates this aesthetic filtering

> We currently provide the following checkpoints:
>
> - `512-base-ema.ckpt`: 550k steps at resolution `256x256` on a subset of LAION-5B filtered for explicit pornographic material, using the LAION-NSFW classifier with `punsafe=0.1` and an **aesthetic score** **`>= 4.5`**. 850k steps at resolution `512x512` on the same dataset with resolution `>= 512x512`.

Aesthetic Ranking

"To create LAION-Aesthetics we trained several lightweight models that predict the rating people gave when they were asked "How much do you like this image on a scale from 1 to 10?". - (Schuhmann 2022)

In order to categorize billions of images according to an aesthetic score, LAION created aesthetic predictor models to facilitate the process. From LAION:

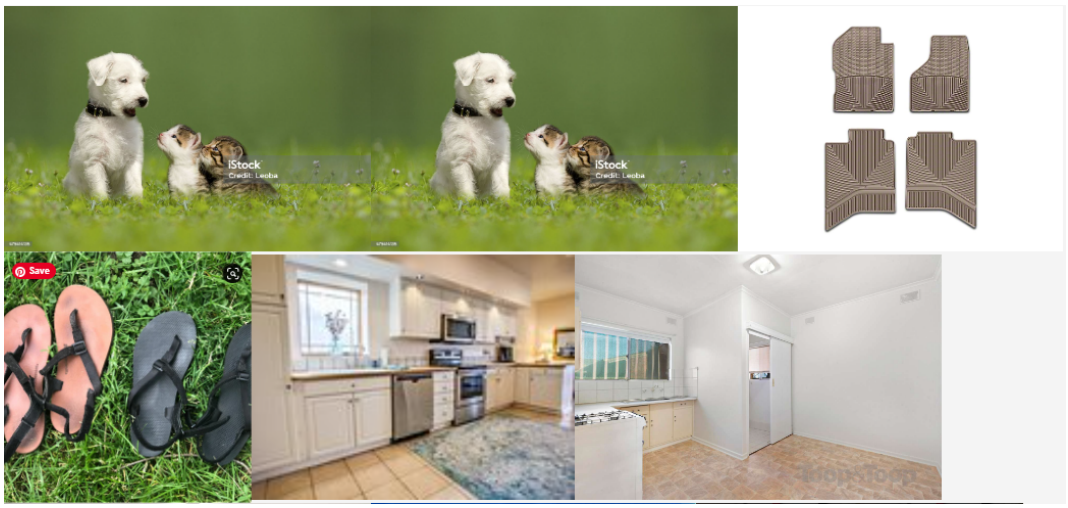
> Using LAION-Aesthetics_Predictor V2, we created the following subsets of the LAION 5B samples with English captions:
>
> 1,2B image-text pairs with predicted aesthetics scores of 4.5 or higher
> 939M image-text pairs with predicted aesthetics scores of 4.75 or higher

600M image-text pairs with predicted aesthetics scores of 5 or higher
12M image-text pairs with predicted aesthetics scores of 6 or higher
3M image-text pairs with predicted aesthetics scores of 6.25 or higher
625K image-text pairs with predicted aesthetics scores of 6.5 or higher
These subsets overlap. 5 fully includes 6 which includes 6.25 and so on. We call the collection of these subsetsLAION-Aesthetics V2.

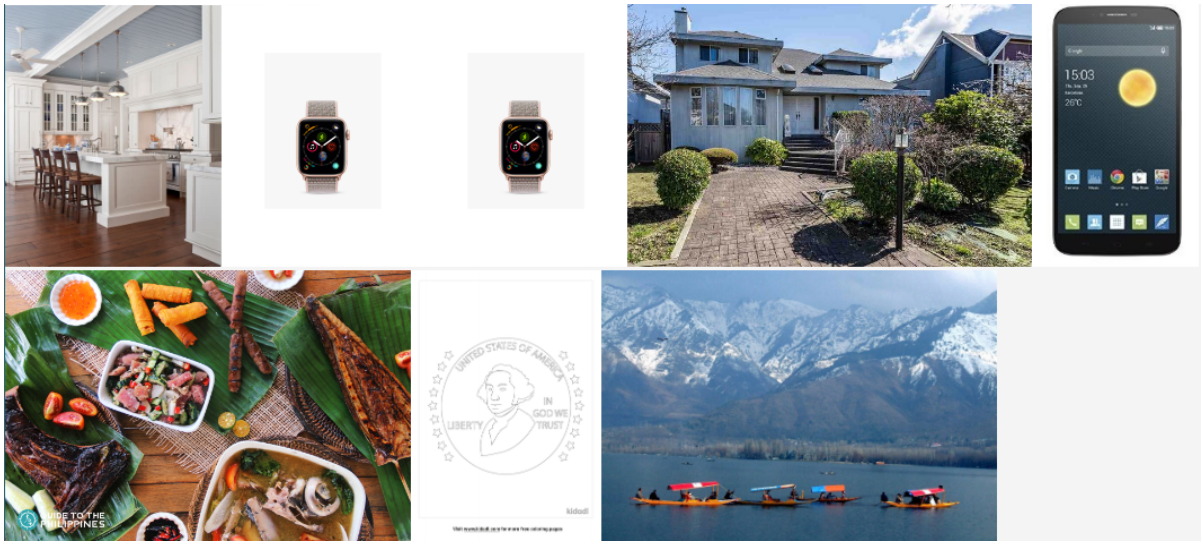
Images that fell below a rank of 4.5 were not included in the creation of aesthetic datasets. From here 5 billion images are reduced to 1.2 billion, with the number of urls dropping to only 12 million images at a score above 6.

Image Samples by Range

4.5



5.0

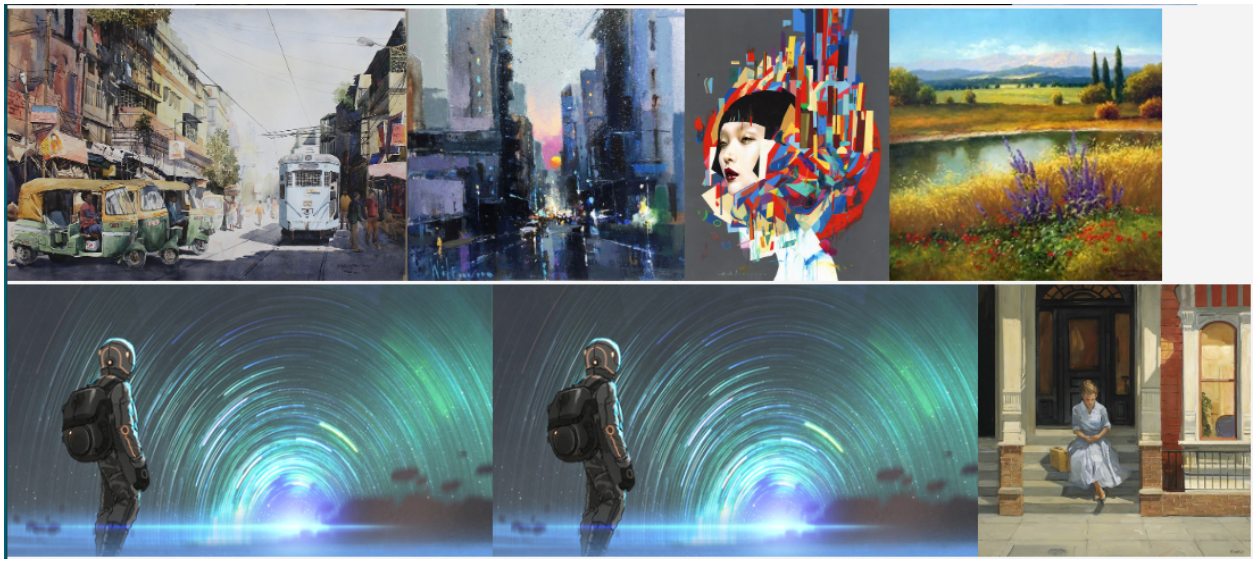


5.5

6.0

6.5

7.0



7+

These examples demonstrate only a small sampling of the images found in their respective aesthetic ranges. https://whatsinlaion.com/ allows filtering by any score contained in the improved_aesthetics_4.5plus dataset.

We do not have examples of images that are scored below 4.5, but we can see as the number increases we are more likely to see work made by professionals. This demonstrates a bias towards sourcing images provided by particular individuals and services, and omitting content that doesn't fall within the criteria of 'aesthetic quality' (the remaining 3.8 billion images unused from LAION5B).

Distribution Ranges

For the creation of https://whatsinlaion.com, the totals were broken down differently than LAION's aesthetic ranges. This is helpful to know for later when we zoom into contributions from particular content providers.

Aesthetic Range	Total # URLs
4.5 - 5.0	766,601,696
5.0 - 5.5	481,529,790
5.5 - 6.0	111,823,637
6.0 - 6.5	11,461,253
6.5 - 7.0	621,877
7.0+	13,683

For completeness, the following are totals for the base LAION datasets.

Dataset	Total # URLs
LAION2B-en	2,322,108,899
LAION2B-multi	2,266,202,925
LAION1B-nolang	1,271,703,630

Content Providers

In order to get a sense of where images in LAION are hosted we needed to parse each url entry for their domain. Meaning for an image located at the following:

https://i.pinimg.com/236x/36/df/20/36df20c0afe1a1826b28f1497d010deb.jpg

We parse the root address of *'i.pinimg.com'* and at that as a datapoint. From this we can see the total number of unique hosts, as well as determine the total number of images each host contributes to a given dataset or distribution range.

LAION-5B contains 22,120,659 million unique hosts.
improved_aesthetics_4.5 plus contains 9,489,747 unique hosts.

Totals

After parsing each entry, we can then determine total contributions of urls to each dataset. The following tables are a (very small) sample of that data.

LAION2B-en Totals

domain	# urls
cdn.shopify.com	134,989,900
i.pinimg.com	85,190,559
i.ebayimg.com	37,013,317
images-na.ssl-images-amazon.com	28,936,548
thumbs.dreamstime.com	24,189,088
www.specsserver.com	23,162,113
i0.wp.com	23,078,056
render.fineartamerica.com	22,986,509
i.ytimg.com	20,182,019
images.slideplayer.com	18,485,777
…	….

improved_aesthetics_4.5plus (total)

domain	# urls
cdn.shopify.com	91,245,549
i.pinimg.com	61,079,061
www.specserver.com	21,651,319
thumbs.dreamstime.com	18,521,722
render.fineartamerica.com	18,093,228
us.123rf.com	17,070,160
i0.wp.com	13,915,292
images-na.ssl-images-amazon.com	12,094,627
…	…

Each Totals page holds results for millions of entries, making a website database one of the better ways to display this information. It is accessible at https://whatsinlaion.com/aesthetics/totals.

I will also upload separate .xlsx documents containing sample data from these totals.

Detail View

In the Appendix that follows I will provide sample data on specific domains. All of the information found here is visible on https://whatsinlaion.com/ when viewing a specific domain.

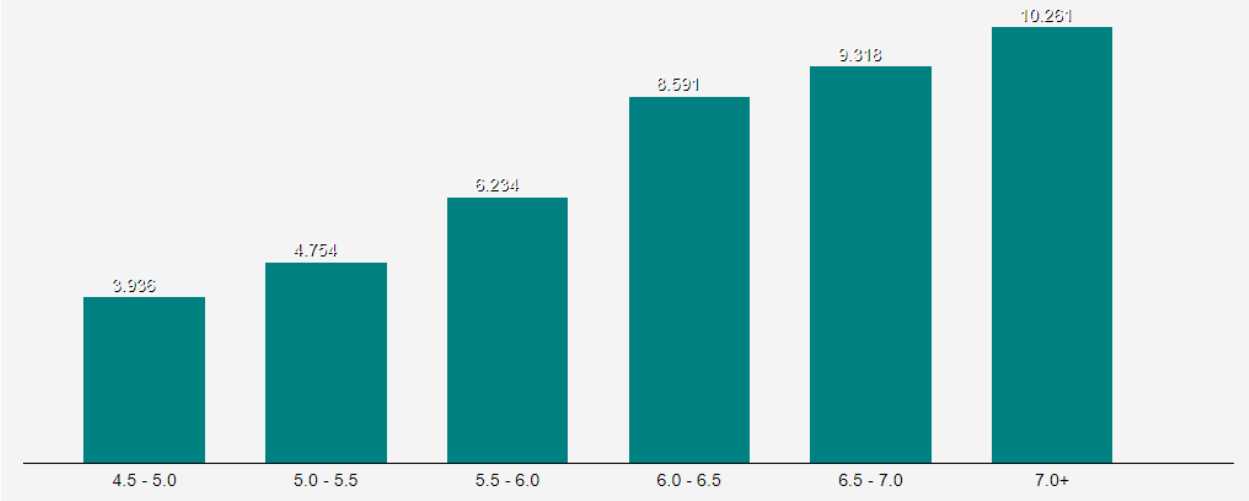
aesthetic contributions

Viewing the info page for any specific domain reveals the totals for each aesthetic range outlined above, as well as their percentage contribution to those ranges.

aesthetic range	# images	% contributions per distribution range
4.5 - 5.0	30,170,394	3.936
5.0 - 5.5	22,893,147	4.754
5.5 - 6.0	6,971,571	6.234
6.0 - 6.5	984,596	8.591
6.5 - 7.0	57,949	9.318
7.0+	1,404	10.261
total	61,079,061	

aesthetic ranges for 'i.pinimg.com'

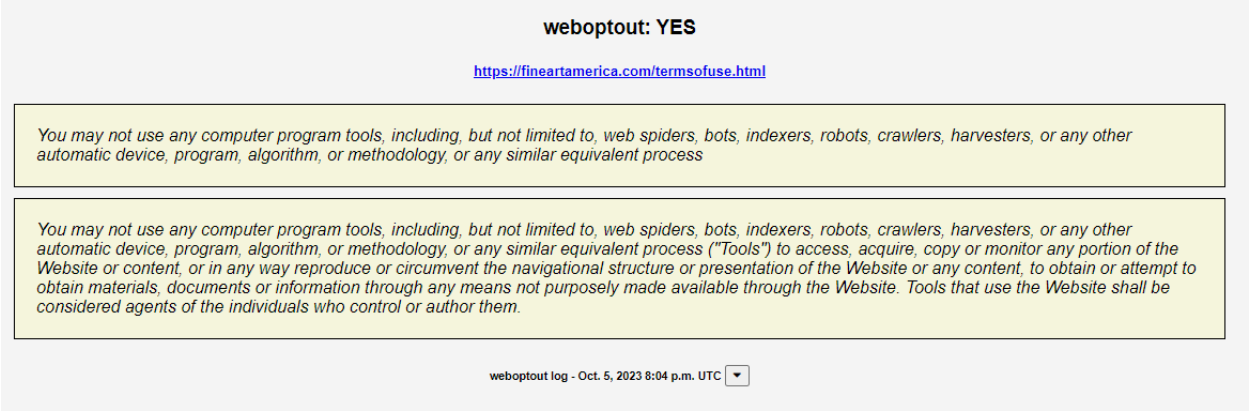
The percentages are accompanied by a graph normalized to the domain's highest contributing percentage. This graph may, at a glance, show which aesthetic range a site's images trend to contribute the most.



% contributions per range for 'i.pinimg.com'

weboptout

For some domains, their website contains Terms of Service language that describes that using crawlers and scrapers is prohibited. "weboptout" is an open source Python library created by Alex J. Champanard that automates checking a site for its Terms of Service, looking for language on whether or not scraping is permitted. (Champandard, n.d.)



Questions

The compilation of this data and creation of https://whatsinlaion.com is to allow researchers to answer some questions for themselves. For any given host found, consider the following:

- What is the intended purpose of the website and what content does it host? Who provides the content?
- What rights are users required to have before submitting content to the host?
- What incentive do platforms have to regulate unlicensed or infringing images being uploaded?
- Is web scraping in violation of the site's Terms of Service?
- Outside of Terms of Service, what methods can content hosts use to prevent the unlicensed use of imagery found on their site?
- What incentive do platforms have to prevent the unlicensed use of imagery found on their site?
- What rights to uploaded material do content providers have? How are they evaluating the provenance of each item submitted to their sites? Or are they not?
- Do site-wide opt-out provisions such as robots.txt truly cover the provenance of scraped data, when it's possible to deep link to images from outside hosts (such as from this site)?

Conclusion

It is my hope that by quantifying difficult to reach information, I have shed more light onto data collection processes used by AI developers. Though the methods used in training Stable Diffusion 1 and 2 are more transparent than most, the amount and sources of those materials are not easily quantified for public consumption.

I invite the Copyright Office to delve further into the information provided, either through the use of https://whatsinlaion.com or by reaching out to me directly.

References

Beaumont, Romain. 2022. “LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS.” LAION. https://laion.ai/blog/laion-5b/.

CC. n.d. “Common Crawl.” Common Crawl - Open Repository of Web Crawl Data. Accessed October 30, 2023. https://commoncrawl.org/.

“cc2dataset.” n.d. github.com/rom1504/cc2dataset. https://github.com/rom1504/cc2dataset.

Champandard, Alex J. n.d. “alexjc/weboptout: Opt-Out tool to check Copyright reservations in a way that even machines can understand.” GitHub. Accessed October 30, 2023. https://github.com/alexjc/weboptout.

“@EMostaque - Jan 19, 2023.” 2023. X/Twitter. https://twitter.com/EMostaque/status/1616130747710029855.

“@EMostaque - Mar 13, 2023.” 2023. X/Twitter. https://x.com/EMostaque/status/1635441014000295938?s=20.

Rombach, Robin, Dominik Lorenz, and Patrick Esser. n.d. “runwayml/stable-diffusion-v1-5 · Hugging Face.” Hugging Face. Accessed October 30, 2023. https://huggingface.co/runwayml/stable-diffusion-v1-5.

Rombach, Robin, Dominik Lorenz, and Patrick Esser. n.d. “stabilityai/stable-diffusion-2-1 · Hugging Face.” Hugging Face. Accessed October 30, 2023. https://huggingface.co/stabilityai/stable-diffusion-2-1.

Schuhmann, Christoph. 2022. “LAION-Aesthetics.” LAION. https://laion.ai/blog/laion-aesthetics/#laion-aesthetics-v2.

“Web crawler.” n.d. Wikipedia. Accessed October 30, 2023. https://en.wikipedia.org/wiki/Web_crawler.

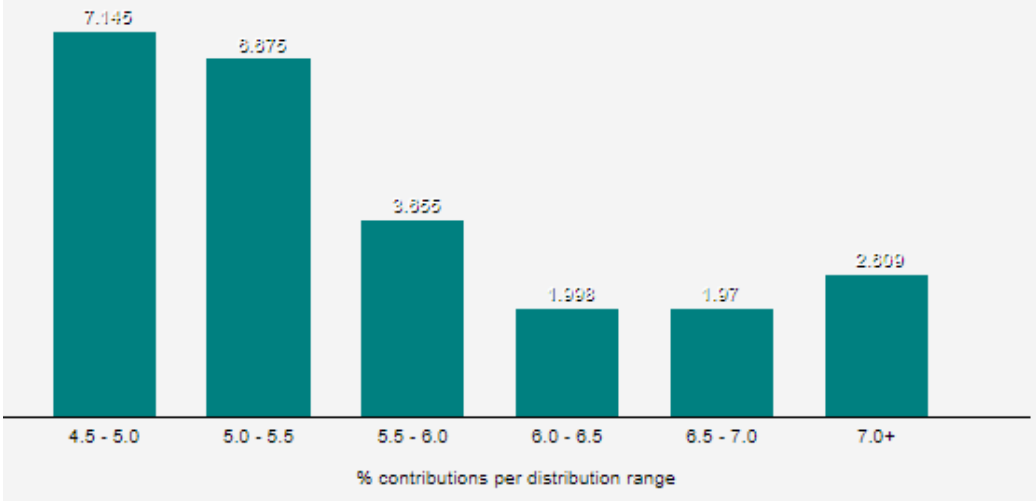
Appendix - Domain Studies

cdn.shopify.com

https://www.shopify.com/legal/terms

You agree not to access the Services or monitor any material or information from the Services using any robot, spider, scraper, or other automated means.

aesthetic range	# images	% per distribution
4.5 - 5.0	54,776,177	7.145
5.0 - 5.5	32,140,795	6.675
5.5 - 6.0	4,086,945	3.655
6.0 - 6.5	229,027	1.998
6.5 - 7.0	12,248	1.97
7.0+	357	2.609
total	91,245,549	

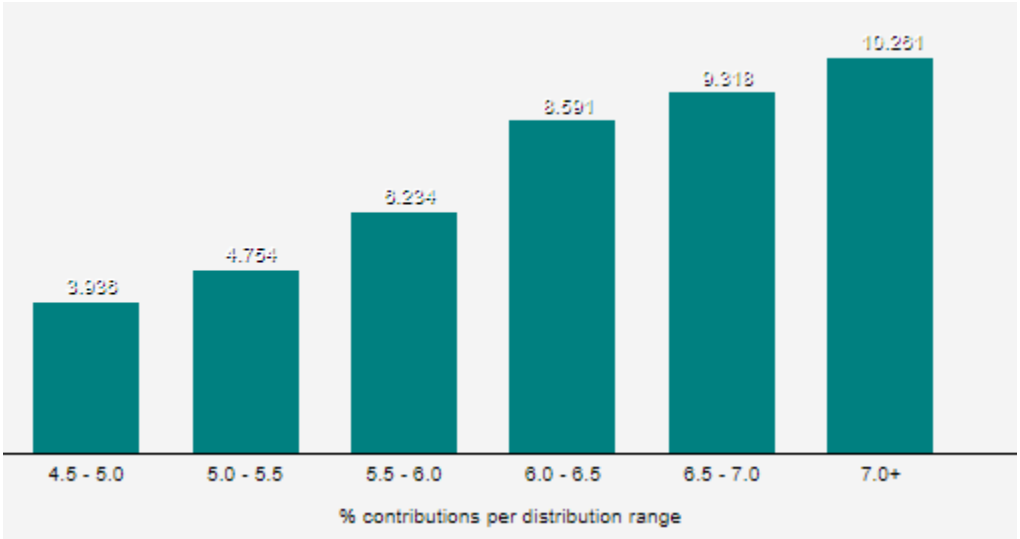


i.pinimg.com

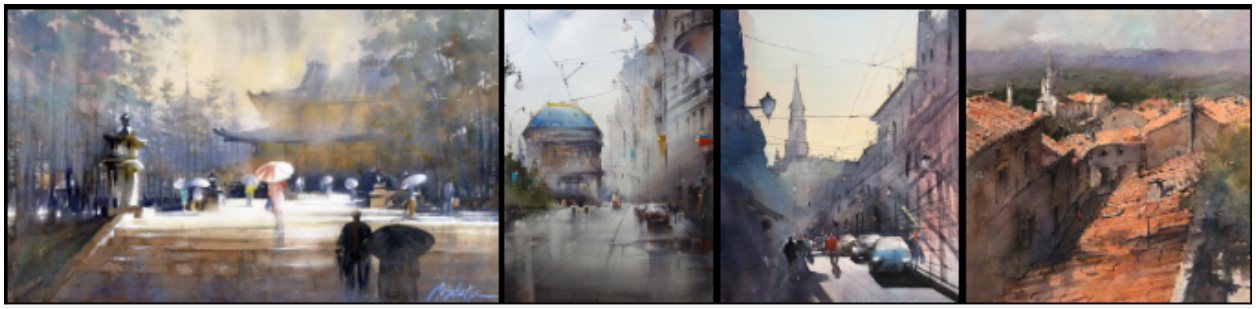
https://policy.pinterest.com/en/terms-of-service

In using Pinterest, you agree not to scrape, collect, search, copy or otherwise access data or content from Pinterest in unauthorized ways, such as by using automated means (without our express prior permission), or access or attempt to access data you do not have permission to access.

aesthetic range	# images	% per distribution
4.5 - 5.0	30,170,394	3.936
5.0 - 5.5	22,893,147	4.754
5.5 - 6.0	6,971,571	6.234
6.0 - 6.5	984,596	8.591
6.5 - 7.0	57,949	9.318
7.0+	1,404	10.261
total	61,079,061	



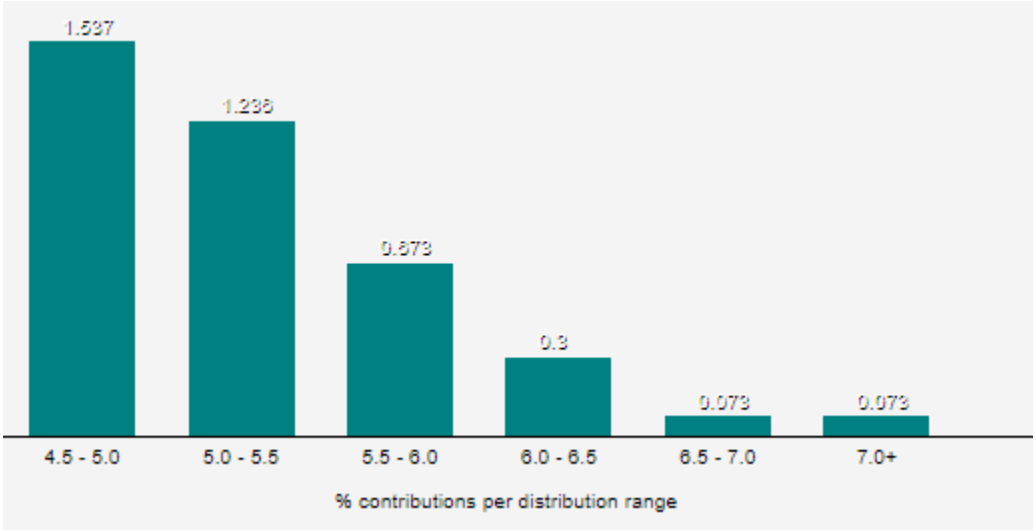
thumbs.dreamstime.com



https://www.dreamstime.com/terms

Additionally, we do not allow the use of automated software or other crawling techniques for searching our web site and/or retrieve Media or related information.

aesthetic range	# images	% per distribution
4.5 - 5.0	11,783,893	1.537
5.0 - 5.5	5,949,993	1.236
5.5 - 6.0	752,976	0.673
6.0 - 6.5	34,399	0.3
6.5 - 7.0	451	0.073
7.0+	10	0.073
total	18,521,722	

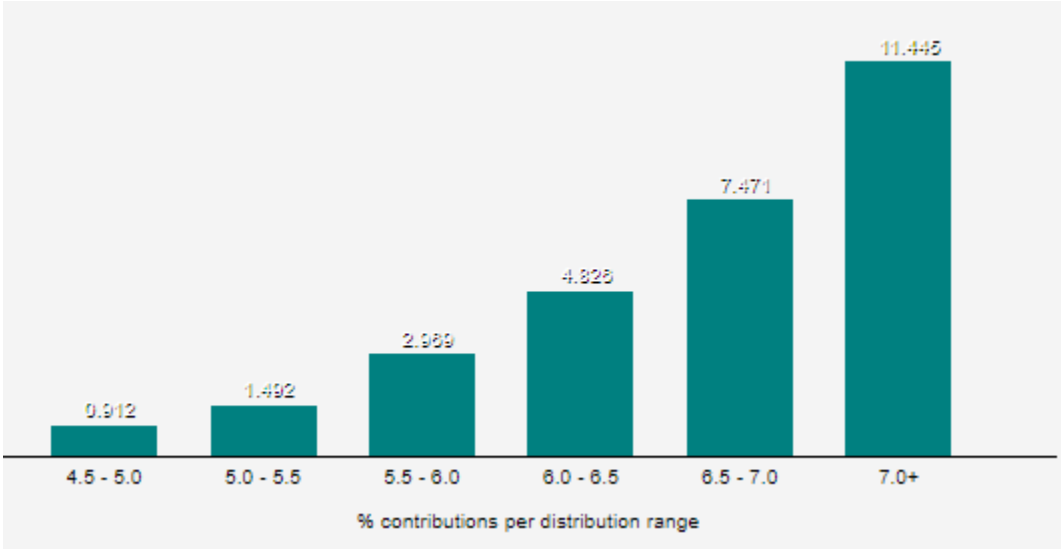


render.fineartamerica.com

https://fineartamerica.com/termsofuse.html

You may not use any computer program tools, including, but not limited to, web spiders, bots, indexers, robots, crawlers, harvesters, or any other automatic device, program, algorithm, or methodology, or any similar equivalent process

aesthetic range	# images	% per distribution
4.5 - 5.0	6,990,183	0.912
5.0 - 5.5	7,182,071	1.492
5.5 - 6.0	3,319,868	2.969
6.0 - 6.5	553,078	4.826
6.5 - 7.0	46,462	7.471
7.0+	18,093,228	11.445
total	18,521,722	

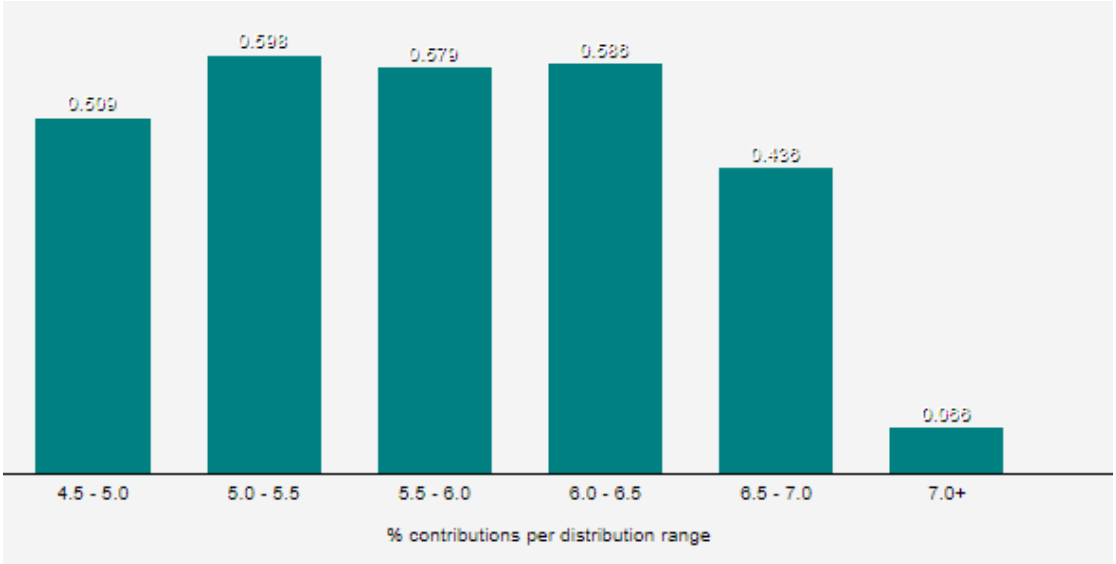


ssl.c.photoshelter.com

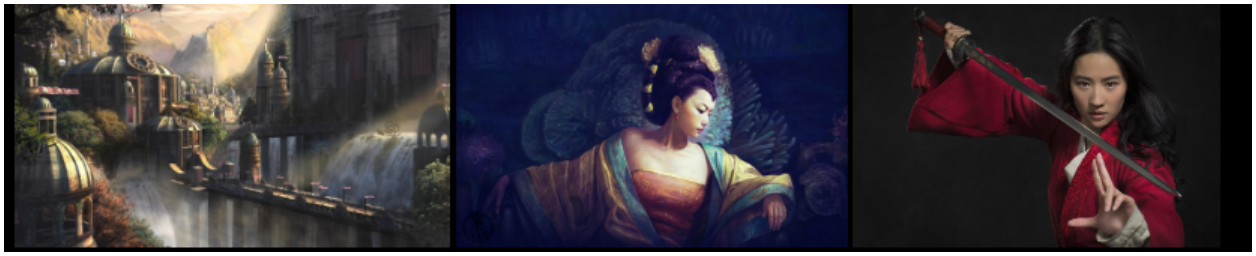
https://company.photoshelter.com/terms/

Rules of Conduct. a. You shall not (and shall not permit any third party to) either (1) take any action or (2) upload, download, post, submit or otherwise distribute or facilitate distribution of any Content on or through the Service, including without limitation any User Content, that: infringes any patent, trademark, trade secret, copyright, right of publicity or other right of any other person or entity

aesthetic range	# images	% per distribution
4.5 - 5.0	3,900,565	0.509
5.0 - 5.5	2,881,721	0.598
5.5 - 6.0	647,369	0.579
6.0 - 6.5	67,169	0.586
6.5 - 7.0	2,711	0.436
7.0+	9	0.066
total	7,499,544	



cdn.wallpapersafari.com



https://wallpapersafari.com/page/terms-of-service/

This web site is intended to be accessed via standard web browser software such as the ones present on our Compatible Browsers section, and similar products via direct interaction by a human. With the exception of publicly accessible RSS feeds provided in XML format, the web site and its associated files are not meant to be accessed via any automated means such as by scripts or bots or automated applications.

aesthetic range	# images	% per distribution
4.5 - 5.0	56,210	0.007
5.0 - 5.5	96,492	0.02
5.5 - 6.0	59,049	0.053
6.0 - 6.5	12,953	0.113
6.5 - 7.0	917	0.147
7.0+	5	0.037
total	225,626