Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The Public Domain 12M (PD12M) dataset was created to provide text-to-image model trainers with a large, high-quality dataset that minimizes copyright risks.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The dataset was created by Jordan Meyer and Nicholas Padgett on behalf of the organization Spawning.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Spawning funded the creation of the dataset. Hosting of the images is provided by Amazon through their AWS Open Data Sponsorship Program.

Any other comments?

None

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance of the dataset is an image believed to be in the public domain, along with a synthetically generated caption describing the image.

How many instances are there in total (of each type, if appropriate)?

The PD12M dataset has 12.4M items, matching the published count of Conceptual Captions 12M in May, 2020. We also release a 3.3M subset of PD12M, matching the published count of the original Conceptual Captions dataset.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

PD12M is a subset of 38M public domain images collected by Spawning in the first quarter of 2024. The full 38M images are available for review at https://Source.Plus. The 12M subset is not a representative sample of the full collection. Our curation process was intended to make the dataset most useful for highly aesthetic T2I model training.

The largest source of divergence from the full collection comes from excluding ~9M document scans. We also excluded materials programmatically identified as NSFW. We used the advanced search features of Source.Plus to manually identify toxic content in the metadata (e.g. ethnophaulisms in the titles or descriptions) and copyrighted content that was misidentified in our sources. We removed duplicated images, keeping the records with the largest dimensions and most complete provenance information. Finally, we sorted the remaining images by the aesthetic ranking of an internal model and selected the 12.4M highest rated images to comprise the PD12M dataset. We further filtered to the top 3.3M images by aesthetic score to create the PD3M dataset.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

The parquet files released on Hugging Face contain URLs pointing to each image. The unprocessed images themselves are hosted separately on AWS, and are maintained by Spawning. Hosting the images specifically for their use as part of the dataset was a necessary step to avoid externalizing the costs of serving the images onto the organizations who originally provided them.

Is there a label or target associated with each instance? If so, please provide a description.

We include synthetic captions for each image alongside the URL in the parquet files. These were generated using Microsoft's Florence-2-large model.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Every image included in the dataset has an associated caption, and can be thought of as complete for the purpose of training a text-conditioned image generation model.

We did not include the extensive metadata that we collected from the source institutions in this dataset, but that metadata can be accessed at Source.Plus. In future releases of PD12M, PD3M, and other datasets derived from Source.Plus, we expect to include additional metadata that will make the datasets more useful for a wider range of tasks.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Relationships between individual instances in the context of PD12M could include grouping by metadata, such as works: painted by the same artist, hosted by the same institution, or created during the same artistic period. The metadata that covers those relationships is not included in this release, but is expected to be shared in future releases and is currently available on Source.Plus for review.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

We do not provide any recommended splits.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

~7M images in PD12M were sourced from the Wikimedia Commons' Public Domain and CC0 categories, which contains a mix of user-created works and institutional datasets. Though the community at Wikimedia commons is rigorous and we took additional measures to filter images from these categories that might have been misclassified, we suspect that some misclassified images remain.

~3M images in PD12M were sourced from OpenGLAM organizations, who, in most cases, assigned the public domain or CC0 license to the works they made available. While exceedingly rare, we did find examples of images incorrectly marked as Public Domain by the publisher themselves. Most OpenGLAM organizations are explicit that they've made their best effort to correctly identify public domain works, but make no guarantees as to the accuracy.

The remaining ~2.5M images in PD12M were sourced from iNaturalist and contributed by individual members of its community. It is possible that some of those contributions were assigned a CC0 license by someone who didn't take the original photograph.

Our dataset maintenance procedures (detailed later in this document) are designed to replace misclassified images quickly when they are discovered.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Due to its size (>30TB of images), the full contents of the dataset are split between an AWS bucket containing only the images, and a Hugging Face dataset containing the captions and links to the images.

The AWS storage is provided by Amazon for 2 years. At the end of that period they reevaluate the dataset's eligibility for their Open Data Sponsorship Program, and Spawning is committed to relocating the images if needed at that time.

The only changes we anticipate making to the dataset are corrections of errors or removal of otherwise problematic images that are discovered after publication. In those instances, we will replace the offending images with their most similar alternative. This ensures that the data remains as reproducible as possible, while acknowledging that no dataset of this size will be without issues that should be addressed upon discovery.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

After extensive manual searching, we have not identified any confidential data in any of the sources.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Yes. Digitized public domain content over-represents Western geographic regions and cultures as well as older content. Despite careful curation, offensive images reflecting these biases will invariably persist.

We performed extensive manual searching for offensive material, and we removed several hundred examples. Given that art, the subject of this dataset, often intends to provoke, the lines of what to remove are not always clear. Our criteria for removal was similar to the framing of this question.

Our flagging system is designed to allow flagging for any reason. Once flagged, images are hidden on the Source.Plus platform. Upon review, images that meet the above criteria will be replaced. Researchers who would like to audit the images that have been removed and the reason why are welcome to contact us.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

The dataset does not directly identify any subpopulations or provide information that could identify individuals. All original image captions were stripped and replaced with synthetic captions to mitigate any risk of personally identifiable details that might have been included in the original image metadata.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

After extensive manual searching, we have not identified any examples of this form of sensitive data for living persons.

Any other comments?

None

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

We originally collected 26.3M images directly from a variety of museums, libraries, scientific and cultural heritage organizations, as well as aggregators of their content, like Europeana. We also sourced 11.3 images from Wikimedia Commons, which contains a mix of user uploads and institutional datasets.

In all cases, we only collected images with metadata indicating a Public Domain Mark or CC0 license.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The sources of the images typically make their metadata available by API access, metadata dumps, or both. One of our primary goals in gathering the data was to minimize the impact on the servers of the hosting institutions. Where available, we gathered metadata from the static files to avoid traversing APIs.

Once the metadata was collected, we downloaded the images from their provided links over the course of two month, using self-imposed rate limits. We store the images on AWS, so that future downloads of the datasets don't impact the providers.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

PD12M and PD3M followed a deterministic filtering process focused on quality, aesthetics and safety.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Data collection was completed by Spawning employees in the course of their regular job duties.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Metadata collection occurred in Q2 of 2024 and the image downloading process continued into 2024-Q3 due to self-imposed rate limiting.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

During the collection period we held frequent discussions with stakeholders and experts that included lawyers, developers, OpenGLAM organizers, and, critically, visual artists.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Data was collected via third parties. Primarily museums, libraries, and other scientific and cultural heritage institutions.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

One of the primary motivations of using only PD and CC0 material was to ensure that living creators have explicitly provided their consent to use their work without any obligation to notify them.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

We limited our collection to include only items that explicitly allow use for any purpose, and otherwise did not contact the individual creators of the images directly. The CC0 license indicates "no rights reserved." Likewise, Public Domain Mark indicates that works are no longer under copyright.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

We checked the source URLs against our Do Not Train (DNT) registry, to respect any preference against AI training. The DNT protects over 2B distinct URLs and none of the 38M image links we originally collected for PD12M were present. We will continue to honor any revocations via our dataset update processes.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No data protection impact analysis has been conducted.

Any other comments?

None

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

We used automated and manual filtering to ensure the overall quality of the dataset. Our curation process was primarily performed through Source.Plus and reduced the 38M images from our initial collection process to 12.4 million for PD12M and to 3.3 million for the PD3M subset.

Semantic Embeddings: We first used CLIP ViT-L/14 to generate embeddings, which served as inputs for downstream curation tasks, including document scan identification, NSFW filtering, and aesthetic scoring.

Document Scan Identification: For the first filtering step, we tagged 8.7M images using an XGBoost model trained to identify document scans using semantic embeddings as features.

Format and Resolution Restrictions: We then imposed a minimum resolution threshold of 256x256 pixels.

Content Filtering: We used LAION's NSFW Detector to exclude works with a score greater than 0.5 (on a scale of 0–1). We also manually reviewed the dataset using semantic search tools to remove instances of non-artistic photographic nudity.

Additionally, we completed a manual check of known ethnophaulisms listed in Wikipedia. For each term, we searched metadata and conducted a semantic search to remove derogatory images.

These steps flagged fewer than 0.05% of the total images collected, demonstrating the high value of limiting our initial image collection to trusted sources.

Deduplication: We represented the dataset's images as nodes in a sparse graph, with links created between images when the cosine distance of their SSCD embeddings was <0.1 (empirically determined). Each subgraph with more than one member was treated as a group of duplicates. For each duplicate group, we selected a canonical image by choosing the item having: 1) a GLAM source, if available, 2) the largest image dimensions, 3) the highest aesthetic score, 4) the largest file size, and 5) the most complete metadata.

Aesthetic Scoring: We assigned each downloaded image an aesthetic score using an XGBoost model trained on an internal dataset of human-ratings. For our final curation step, to match the original size of CC12M (12.4M), we excluded images from the bottom TK% of aesthetic scores. To match the original size of CC3M (3.3M), we excluded the bottom TK% of images by aesthetic scores.

Manual Spot-checks: Throughout the curation process, we performed spot-checks on random samples to verify the effectiveness of the automated filtering. We estimate that our spot-checks covered ~0.01% of the dataset, with a focus on edge cases and copyright misclassifications.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw data that we collected is available for public review on the Source.Plus website. If an image was identified as copyrighted after we downloaded it, we removed our copy of the image and hid the metadata from public view. The metadata for these and other images that we manually flagged are available for review upon request to the authors.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

The cleaning, preprocessing, and curation steps were conducted manually using open-source tools and on the Source.Plus platform.

Any other comments?

None

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Subsets of the dataset were used to train classifier models that we used to filter the document scans, tag concepts, and provide aesthetic ratings. The full PD12M dataset in its current form has not been used to train a text-to-image model as of its release (Oct 2024).

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

NA

What (other) tasks could the dataset be used for?

In its current form, it is most useful for text-to-image modeling. Image-to-text models could be useful, but would be limited by the accuracy of our synthetic captions. With the additional metadata planned for future releases, class-based image generation and classifier models for artistic categorization (eg medium, period, artist) will also be possible.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

The dataset's focus on public domain materials creates an emphasis on Western art and art styles that future users should take into consideration.

While we believe that this is the most comprehensive attempt to create a public domain only image-text dataset to date, the sources of the original images, and by extension, Spawning, cannot fully guarantee that no copyrighted material appears in the dataset. Users should ensure that they are using the most recent version of PD12M, as we plan to address any instances of copyrighted material found in the dataset as they arise.

Are there tasks for which the dataset should not be used? If so, please provide a description.

We chose a permissive license to allow for any use of the dataset, echoing the spirit of the items within it.

Any other comments?

None

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset may be distributed freely.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

Parquet files with the captions and image URLs are distributed as a Hugging Face dataset. The images themselves are hosted on AWS and can be downloaded using the image URLs enumerated in the parquet files.

When will the dataset be distributed?

The datasets are available as of Oct 24, 2024.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Community Data License Agreement - Permissive
https://cdla.dev/permissive-2-0/

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

We are not aware of any other restrictions on any of the items within the dataset. If any are found, the items will be immediately hidden on Source.Plus and removed from the Hugging Face and AWS storage within one week of discovery.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

We are not aware of any export controls or regulatory restrictions on any of the items within the dataset or of the dataset itself.

Any other comments?
None

Maintenance

Who will be supporting/hosting/maintaining the dataset?
Spawning will support and maintain the dataset. The metadata and images are hosted by Hugging Face and AWS respectively.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
jordan@spawning.ai

Is there an erratum? If so, please provide a link or other access point.
The dataset metadata is mirrored in a Source.Plus collection. Post-publication changes will be visible through the change logs of the collection.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?
We plan to continue to add columns to the dataset as we review the licensing terms of the metadata. We do not plan to add additional rows to these two datasets, though we do plan to release additional PDxM datasets as we grow the Source.Plus collection.

We also expect to update the dataset when problematic images are discovered. These items will be replaced with the most similar image available on Source.Plus that is not already included in the dataset. This process is intended to maximize the dataset reproducibility while encouraging community-led auditing.

Source.Plus provides keyword, semantic, faceted, and reverse image search capabilities to enable dataset exploration and auditing. Users can flag items for any reason, including copyright, bias, and privacy concerns, and record their justification in a free-text field. When an item is flagged, it becomes immediately invisible to users of Source.Plus pending review. Reviewed items are either confirmed for replacement or restored. We expect to update the dataset files in the AWS and Hugging Face repositories within a week of confirmed replacements.

All changes will be communicated via a changelog that will be linked at the Source.Plus collection and on the Hugging Face Datasets page.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.
There are no automatic limits on the data retention. Requests by any individual who created or is depicted in an image, regardless of its license, will be removed and replaced according to the process from the previous answer.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.
This is the first release of the PD12M and PD3M datasets. All versions of the metadata will remain available as columns are added or changes are made.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.
We plan to release new, larger datasets as we incorporate additional sources. We are actively searching for additional repositories of Public Domain images that we haven't included, and we welcome suggestions. We are particularly interested in sources with diverse representation and more recent imagery than is typically found in the Public Domain.

Source.Plus also allows for public sharing of CC0 data uploaded by the creator. We manually verify that the contributor has the rights to apply the license to the images. Once verified, the images are displayed in the public collections and will be available for the replacement and curation processes of existing and future datasets.

Any other comments?
None