Segment Anything 1B Dataset Card

We provide a Dataset Card for SA-1B, following Datasheets for Datasets[1] in the subsequent list of questions and answers.

Motivation

1. *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 contributions of our dataset to the vision community are fourfold: (1) We release a dataset of 11M images and 1.1B masks, by far the largest segmentation dataset to date. (2) The dataset we release is privacy protecting: we have blurred faces and license plates in all images. (3) The dataset is licensed under a broad set of terms of use which can be found at https://ai.facebook.com/datasets/segment-anything. (4) The data is more geographically diverse than its predecessors, and we hope it will bring the community one step closer to creating fairer and more equitable models.
2. *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 the FAIR team of Meta AI. The underlying images were collected and licensed from a third party photo company.
3. *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.* Meta AI funded the creation of the dataset.
4. *Any other comments?* No.

Composition

1. *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.* All of the instances in the dataset are photos. The photos vary in subject matter; common themes of the photo include: locations, objects, scenes. All of the photos are distinct, however there are some sets of photos that were taken of the same subject matter.
2. *How many instances are there in total (of each type, if appropriate)?* There are 11 million images.
3. *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).* The dataset is composed of images licensed from a photo provider. The dataset contains all instances licensed. The images are photos, ***i.e.*** not artwork, although there are a few exceptions. The dataset includes all generated masks for each image in the dataset. We withheld ~2k randomly selected images for testing purposes.
4. *What data does each instance consist of? "Raw" data (*e.g.*, unprocessed text or images) or features? In either case, please provide a description.* Each instance in the dataset is an image. The images were processed to blur faces and license plates to protect the identities of those in the image.
5. *Is there a label or target associated with each instance? If so, please provide a description.* Each image is annotated with masks. There are no categories or text associated with the masks. The average image has ~100 masks, and there are ~1.1B masks in total.
6. *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.* Yes. Each image is accompanied by a short caption that describes the content and place of the photo in

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM, 2021.*

a free form text. Per our agreement with the photo provider we are not allowed to release these captions. However, we use them in our paper to analyze the geographical distribution of the dataset.

7. *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.* No, there are no known relationships between instances in the dataset.

8. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.* *Errors:* The masks are generated by a segmentation model, so there may be errors or inconsistencies in the masks. *Redundancies:* While no two images are the same, there are instances of images of the same subject taken close together in time.

9. *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.* The dataset is self-contained.

10. *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.* No.

11. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.* We have two safety measures to prevent objectionable content: (1) Photos are licensed from a photo provider and had to meet the terms of service of the photo provider. We requested that all objectionable content be filtered from the images we licensed. (2) If a user observes objectionable image(s) in the dataset, we invite them to report the image(s) at segment-anything@meta.com for removal. Despite the measures taken, we observe that a small portion of images contains scenes of protests or other gatherings that focus on a diverse spectrum of religious beliefs or political opinions that may be offensive. We were not able to produce a filtering strategy that removes all such images and rely on users to report this type of content.

12. *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.* The dataset does not identify any subpopulations of the people in the photos.

13. *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.* No. Images were subjected to a face blurring model to remove any personally identifiable information. If a user observes any anonymization issue, we invite them to report the issue and the image id(s) at segment-anything@meta.com.

14. *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.* The dataset contains scenes of protests, or other gatherings that may suggest religious beliefs, political opinions or union memberships. However, the faces of all people in the dataset have been anonymized via facial blurring, so it is not possible to identify any person in the dataset.

15. *Any other comments?* No.

Collection Process

1. *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.* The released masks associated with each image were automatically inferred by our segmentation model, SAM. The masks that were collected using model-assisted manual annotation will not be released. Quality was validated as described in the Segmenting Anything paper[2].

[2] `https://ai.facebook.com/research/publications/segment-anything`

2. *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 images in the dataset are licensed from an image provider. They are all photos taken by photographers with different cameras.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*, deterministic, probabilistic with specific sampling probabilities)?* We withheld ~2k randomly selected images for testing purposes. The rest of the licensed images are included in the dataset.
4. *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)?* The released masks were automatically inferred by SAM. For details on our model-assisted manual annotation process see our Data Annotation Card in the Segmenting Anything paper. Note these masks will not be released.
5. *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.* The licensed photos vary in their date taken over a wide range of years up to 2022.
6. *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. If the dataset does not relate to people, you may skip the remaining questions in this section.* We underwent an internal privacy review to evaluate and determine how to mitigate any potential risks with respect to the privacy of people in the photos. Blurring faces and license plates protects the privacy of the people in the photos.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (*e.g.*, websites)?* We licensed the data from a third party photo provider.
8. *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.* The images are licensed from a third party who provided appropriate representations regarding the collection of any notices and consents as required from individuals. In addition, all identifiable information (*e.g*. faces, license plates) was blurred. Under the terms of the dataset license it is prohibited to attempt to identify or associate an image with a particular individual.
9. *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.* The images are licensed from a third party who provided appropriate representations regarding the collection of any notices and consents as required from individuals. In addition, all identifiable information (*e.g*. faces, license plates) was blurred from all images. For avoidance of doubt, under the terms of the dataset license it is prohibited to attempt to identify or associate an image with a particular individual.
10. *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 invite users to report at segment-anything@meta.com for image(s) removal.
11. *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.* To eliminate any potential impact on people whose photos are included in the dataset, identifiable information (faces, license plates) has been blurred.
12. *Any other comments?* No.

Preprocessing/cleaning/labeling

1. *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 resized the high-resolution licensed images such that the shorter side is 1500 pixels and only processed the images to remove any identifiable and personal information from the photos (faces, license plates).
2. *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.* No, as we removed the data for safety reasons and to respect privacy, we do not release the unaltered photos.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.* We used the RetinaFace [3] to detect faces in the images. The model used to blur license plates has not been made public.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.* The dataset was used to train our segmentation model, SAM.
2. *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.* No. However, all users of the dataset must cite it, so its use is trackable via citation explorers.
3. *What (other) tasks could the dataset be used for?* We intend the dataset to be a large-scale segmentation dataset. However, we invite the research community to gather additional annotations for the dataset.
4. *Are there tasks for which the dataset should not be used? If so, please provide a description.* Full terms of use for the dataset including prohibited use cases can be found at https://ai.facebook.com/datasets/segment-anything.
5. *Any other comments?* No.

Distribution

1. *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 will be available for the research community.
2. *How will the dataset will be distributed (*e.g.*, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?* The dataset is available at https://ai.facebook.com/datasets/segment-anything.
3. *When will the dataset be distributed?* The dataset will be released in 2023.
4. *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.* Yes. The license agreement and terms of use for the dataset can be found at https://ai.facebook.com/datasets/segment-anything. Users must agree to the terms of use before downloading or using the dataset.
5. *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.* Full terms of use and restrictions on use of the SA-1B dataset can be found at https://ai.facebook.com/datasets/segment-anything.
6. *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.* The license and restrictions on use of the SA-1B dataset can be found at https://ai.facebook.com/datasets/segment-anything.
7. *Any other comments?* No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?* The dataset will be hosted at https://ai.facebook.com/datasets/segment-anything and maintained by Meta AI.
2. *How can the owner/curator/manager of the dataset be contacted (*e.g.*, email address)?* Please email segment-anything@meta.com.
3. *Is there an erratum? If so, please provide a link or other access point.* No.
4. *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)?* To aid reproducibility of research using SA-1B, the only updates will be to remove reported images.

[3] RetinaFace: `https://github.com/serengil/retinaface`
Sefik Ilkin Serengil and Alper Ozpinar. LightFace: A hybrid deep face recognition framework. *ASYU, 2020.*
Sefik Ilkin Serengil and Alper Ozpinar. HyperExtended LightFace: A facial attribute analysis framework. *ICEET, 2021.*

5. *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 limits on data retention. We took measures to remove personally identifiable information from any images of people. Users may report content for potential removal here: segment-anything@meta.com.

6. *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.* No, as the only updates will be to remove potentially harmful content, we will not keep older versions with the content.

7. *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 encourage users to gather further annotations for SA-1B. Any users who generate annotations will be liable for hosting and distributing their annotations.

8. *Any other comments?* No.