Stanford | Internet Observatory
Cyber Policy Center

Identifying and Eliminating CSAM in Generative ML Training Data and Models

David Thiel
Stanford Internet Observatory
December 20, 2023

Contents

1 Introduction **2**
- 1.1 Development of LAION-5B 2
- 1.2 Structure of LAION-5B 3
- 1.3 Use by subsequent models 3

2 Detecting CSAM in LAION-5B: approaches and limitations **4**

3 Methodology **5**
- 3.1 Validation 6
- 3.2 Cryptographic hashes for missing content 7
- 3.3 Leveraging KNN and ML classification for novel CSAM detection . 8

4 Summary of results **8**

5 Overall findings and recommendations **11**
- 5.1 Removing material 11
- 5.2 Implications for training set compilation 12
- 5.3 Alterations to the model training process 12
- 5.4 Content hosting platforms 13
- 5.5 Model and dataset hosting platforms 13

6 Safety and ethical considerations **14**

7 Conclusion **14**

1 Introduction

Machine learning models that generate visual images are trained on a small number of datasets of images. Many older models, for example, were trained on the manually labeled ImageNet[1] corpus, which features 14 million images spanning all types of objects. However, more recent models, such as Stable Diffusion, were trained on the billions of scraped images in the LAION-5B[2] dataset. This dataset, being fed by essentially unguided crawling, includes a significant amount of explicit material.

While our previous work[3] has indicated that generative ML models can and do produce Child Sexual Abuse Material (CSAM), that work assumed that the models were able to produce CSAM by combining two "concepts," such as child and explicit act, rather than the models understanding CSAM due to being trained on CSAM itself. With the goal of quantifying the degree to which CSAM is present in the training dataset as well as eliminating it from both LAION-5B and derivative datasets, we use various complementary techniques to identify potential CSAM in the dataset: perceptual hash-based detection,[4] cryptographic hash-based detection, and k-nearest neighbors analysis leveraging the image embeddings in the dataset itself. Through this process, we identified 3,226 dataset entries of suspected CSAM, much of which was confirmed as CSAM by third parties.

1.1 Development of LAION-5B

The LAION-5B dataset is derived from a broad cross-section of the web, and has been used to train various visual generative machine learning models. This dataset was built by taking a snapshot of the Common Crawl[5] repository, downloading images referenced in the HTML, reading the "alt" attributes of the images and using CLIP[6] interrogation to discard images that did not sufficiently match the captions. It supplants an earlier dataset, LAION-400M, which was built using a similar process.[7]

LAION-5B is split into several subsets: LAION-2B-en, which contains 2.32 billion entries with descriptions in English, LAION-2B-multi which contains 2.26 billion entries with text descriptions for other languages, and LAION-1B-nolang, which contains 1.27 billion entries where the language of the texts could not be detected.

1. Jia Deng et al., "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–55, https://doi.org/10.1109/CVPR.2009.5206848.

2. Christoph Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022), https://openreview.net/forum?id=M3Y74vmsMcY.

3. David Thiel, Melissa Stroebel, and Rebecca Portnoff, "Generative ML and CSAM: Implications and Mitigations," *Stanford Digital Repository*, https://doi.org/10.25740/jv206yg3793.

4. Hany Farid, "An Overview of Perceptual Hashing," *Journal of Online Trust and Safety* 1, no. 1 (October 2021), https://doi.org/10.54501/jots.v1i1.24.

5. *Common Crawl - Open Repository of Web Crawl Data*, accessed September 15, 2023, https://commoncrawl.org.

6. Alec Radford et al., *Learning Transferable Visual Models From Natural Language Supervision*, 2021, arXiv: 2103.00020 [cs.CV].

7. Christoph Schuhmann et al., *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*, 2021, arXiv: 2111.02114 [cs.CV].

Current open-source machine learning models were primarily trained on LAION-2B-en, with additional training on images gauged to be aesthetically pleasing. However, future models are likely to be trained on models which combine LAION-2B-en and LAION-2B-multi with image descriptions translated to English.[8]

1.2 Structure of LAION-5B

In order to minimize size, liability and copyright issues, the LAION datasets do not contain the actual referenced images themselves; rather, they are metadata about the images, with a URL to the original image at the time of initial collection. The fields relevant to this analysis are as follows:

Table 1: Relevant fields from the LAION datasets.

Field	Description
hash	the image's identifier
URL	URL of the original image
TEXT	The extracted description of the image
LANGUAGE	For the -multi and -nolang datasets, the detected language
ENG_TEXT	The text translated to English, if applicable
punsafe	The detected "unsafe" probability for the image

LAION also distributes image embeddings[9] for each image in the dataset to help calculate image similarity, in versions calculated with the L/14 and H/14 OpenCLIP models.[10] The use of image embeddings in this analysis is discussed further in Section 3.

1.3 Use by subsequent models

While the developers of LAION-5B did attempt to classify whether content was sexually explicit,[11] as well as to detect some degree of underage explicit content,[12] the most popular resultant models (namely Stable Diffusion 1.5) were ultimately trained on a wide array of content, both explicit and otherwise. In the subsequent 2.0 version of Stable Diffusion, results with an "unsafe" value higher than 0.1 were filtered out,[13] resulting in a substantial lack of explicit material in the

8. CLIP was also trained on English text, though this approach may change somewhat with the advent of multilingual CLIP; see Fredrik Carlsson et al., "Cross-lingual and Multilingual CLIP," in *Proceedings of the Language Resources and Evaluation Conference* (Marseille, France: European Language Resources Association, June 2022), 6848–54, https://aclanthology.org/2022.lrec-1.739.

9. Romain Beaumont, "Large Scale OpenCLIP: L/14, H/14 and G/14 Trained on LAION-2B," September 15, 2022, https://laion.ai/blog/large-openclip.

10. LAION, *LAION eV*, Hugging Face repository, accessed October 2, 2023, https://huggingface.co/laion.

11. Christoph Schuhmann, *CLIP-based-NSFW-Detector*, GitHub repository, accessed September 15, 2023, https://github.com/LAION-AI/CLIP-based-NSFW-Detector.

12. Christoph Schuhmann, *crawlingathome*, GitHub repository, accessed September 15, 2023, https://github.com/christophschuhmann/crawlingathome-worker/blob/master/crawlingathome.py#L475.

13. Stability AI, *Stable Diffusion v2 Model Card*, November 23, 2022, https://huggingface.co/stabilityai/stable-diffusion-2.

training set—hence making it difficult to generate explicit content. This was met with widespread dissatisfaction from the community,[14] resulting in Stable Diffusion 1.5 remaining the most popular model for generating explicit imagery.[15] Subsequently, Stability AI released Stable Diffusion 2.1, which took 2.0 as a base and further trained it on both "safe" and moderately "unsafe" material.[16]

LAION datasets have also been used to train other models, such as Google's Imagen, which was trained on a combination of internal datasets and LAION-400M.[17] Notably, during an audit of the LAION-400M, Imagen's developers found "a wide range of inappropriate content including pornographic imagery, racist slurs, and harmful social stereotypes",[18] and deemed it unfit for public use.

2 Detecting CSAM in LAION-5B: approaches and limitations

Starting in September 2023, researchers at the Stanford Internet Observatory undertook a series of investigations to determine the degree to which CSAM is present in a given training set. We pursued several approaches, each with their own limitations. Our preliminary approach was based on the LAION-2B-multi file list with image descriptions translated[19] to English. Images with a high "unsafe" value[20] whose descriptions might indicate CSAM were identified, and the original URLs of those images were then passed to PhotoDNA. While this approach identified some instances of known CSAM, it suffered from several limitations:

Translation quality: Running all non-English captions through high-quality translation models such as those offered by Google or DeepL would likely have been cost-prohibitive, both in terms of money and time. As a result, the "many to many" M2M-100 translation model[21] was used, even though text was only being translated from other languages into English. This appears to have resulted in mis-translations ranging from minor to wildly divergent; in some cases, the translation model would get "stuck" and simply repeat the same word over and over.

Search term comprehensiveness: For obvious reasons, search terms used to find CSAM are not widely distributed; this is of course compounded by the presence of dozens of languages in the dataset, many of which may use native language terms or slang that translate poorly. As an example, even

14. James Vincent, "Stable Diffusion made copying artists and generating porn harder and users are mad," *The Verge*, November 24, 2022, https://www.theverge.com/2022/11/24/23476622/ai-image-generator-stable-diffusion-version-2-nsfw-artists-data-changes.
15. Thiel, Stroebel, and Portnoff, "Generative ML and CSAM: Implications and Mitigations."
16. Stability AI, *Stable Diffusion v2-1 Model Card*, December 7, 2022, https://huggingface.co/stabilityai/stable-diffusion-2-1.
17. Chitwan Saharia et al., *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*, 2022, arXiv: 2205.11487 [cs.CV].
18. Saharia et al., 9.
19. Marianna Nezhurina et al., *LAION Translated: 3B Captions Translated to English from LAION5B*, September 15, 2022, https://laion.ai/blog/laion-translated/.
20. As gauged by Schuhmann, *CLIP-based-NSFW-Detector*.
21. Angela Fan et al., *Beyond English-Centric Multilingual Machine Translation*, 2020, arXiv: 2010.11125 [cs.CL].

a commonly used term such as "loli"[22] in Japanese (ロリ) was frequently translated as the name "Lori", or occasionally the word "LOL".

Image label accuracy: From initial testing, we found most positive PhotoDNA hits were not labeled with text that was strongly indicative of CSAM; i.e., actual CSAM entries in the dataset may have generic-sounding labels while explicit material depicting adults may commonly have ambiguous indicators of youth (teen, schoolgirl, twink, etc). The text descriptions for the majority of initial PhotoDNA hits used generic captions that could apply to either legal or illegal material; therefore we conclude that at least for English language material, text descriptions are of limited utility for identifying CSAM.[23]

Due to these issues, keyword-based analysis was abandoned for subsequent larger-scale analysis. This left us with the remaining limitations:

Dead links: LAION datasets do not include the actual images; instead, they include a link to the original image on the site from which it was scraped. Given that multiple years have elapsed between the time the content was scraped and processed, a large percentage of the URLs passed to PhotoDNA (≈30%) were reported as no longer being active. They may, however, have been used to train models before they were removed from their original URLs, and some likely continue to reside in versions of the datasets retrieved at earlier dates.

Null unsafe values: Approximately 6% of images in the dataset had no unsafe value, for reasons which are unclear.

Comprehensiveness of PhotoDNA: PhotoDNA is naturally limited to only known instances of CSAM, and has certain areas upon which it performs poorly. For example, based on keywords, significant amounts of illustrated cartoons depicting CSAM appear to be present in the dataset, but none of these resulted in PhotoDNA matches.

3 Methodology

Given the identified limitations, we chose to focus our analysis on any entry considered by LAION's safety classifier as "unsafe" at the highest level of confidence (> 0.995), regardless of keywords. These URLs would then be sent to PhotoDNA to detect if any known CSAM was present, and matches would be sent to the Project Arachnid Shield API to have the results validated by Canadian Centre for Child Protection (C3P).[24] Once instances of CSAM were verified, we would use their

22. A Japanese term derived from "lolita" commonly used to denote the presence of either child characters or performers, though it is also used to refer to adult performers that appear or pose as being very young.

23. This may be substantially different in languages of countries where CSAM and CSAM-adjacent material are more tolerated; however, due to translation quality issues, this would need to be assessed on a per-language basis with native speakers and term lists.

24. https://www.protectchildren.ca

image embeddings[25] to run k-nearest neighbors (KNN) queries[26] to find related images in the dataset.

After initial validation, we also implemented industry MD5 hash sets provided by the National Center for Missing and Exploited Children (NCMEC)[27] and a CSAM classifier provided by Thorn[28] to discover additional content not detectable via PhotoDNA. A summary of our full methodology is illustrated in Figure 1.

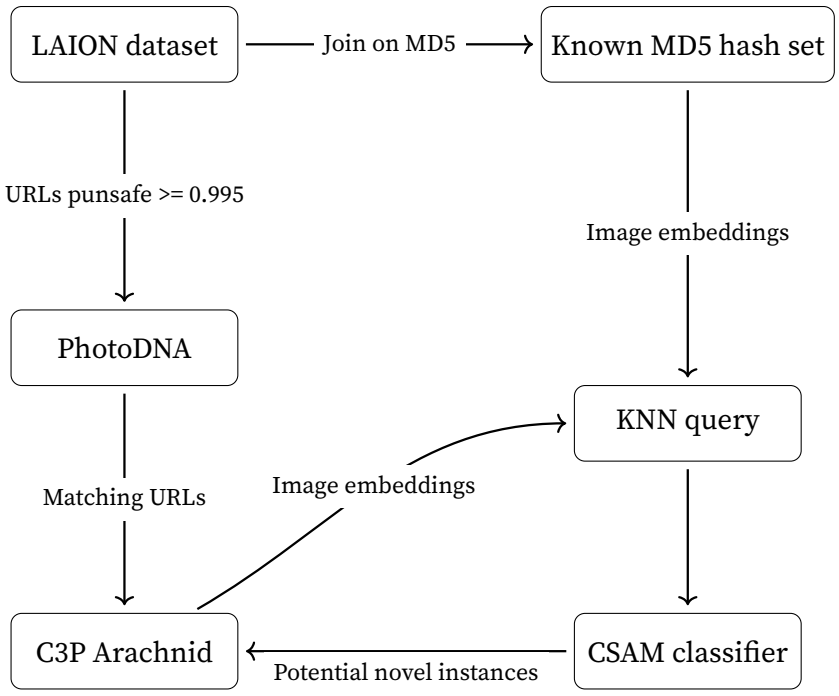

Figure 1: A simplified illustration of the overall methodology used to identify CSAM in the LAION datasets. URLs from the LAION datasets are sent to PhotoDNA; any matches are then sent to the C3P for verification, via the Arachnid API. The image embeddings of verified matches are then used to perform k-nearest neighbors queries to identify additional potential novel instances of CSAM. This process can then be repeated. Separately, MD5 hashes of images in the LAION datasets are compared to known CSAM MD5 hashes, with the image embeddings of matches used to seed further KNN queries.

3.1 Validation

To validate our methodology, we performed a linear assessment of all URLs meeting this cutoff in the first segment of the LAION-2B-multi dataset, resulting in 27 positive matches. We then took these first 27 positive PhotoDNA hits and submitted the URLs to the C3P. Once CSAM was positively identified, we then took the 27 positive matches and performed KNN queries for each. This resulted

25. Romain Beaumont, "Image embeddings," July 20, 2020, https://rom1504.medium.com/image-embeddings-ed1b194d113e.
26. As implemented by the Clip Retrieval package; see Beaumont (*Clip Retrieval: Easily compute clip embeddings and build a clip retrieval system with them*).
27. https://missingkids.org
28. https://www.thorn.org

in 2,773 additional candidates for inspection by PhotoDNA. 1,954 of these images were still live and retrievable by PhotoDNA.

Out of these, 88 additional PhotoDNA hits were found; 43 of them were unique instances (as determined by the PhotoDNA unique match ID), with some images duplicated up to 8 times. Duplication itself is of concern, as the more repetitions of an image in a training set, the more likely it becomes that a model trained on it will produce output highly resembling the repeated training data.[29] This method also allowed us to identify instances of CSAM that had an "unsafe" value below the threshold used to feed the linear search (or a null value).

This confirmed that a) CSAM was identifiably present in the dataset and b) that using KNN could both rapidly identify additional PhotoDNA hits of varying "unsafe" levels and potentially identify novel CSAM.

False positives and what is in PhotoDNA

The issue of "true positives" or "false positives" is somewhat difficult to classify from mere presence in the PhotoDNA database. For example, a video of CSAM is often broken down into individual frames, with perceptual hashes generated for each of these frames. Each frame may not be legally CSAM, but it comes from a video containing CSAM, therefore making its detection important for online platforms. Different PhotoDNA hash databases also contain material that is age-ambiguous, with different participants in the system having different rules for inclusion.

3.2 Cryptographic hashes for missing content

While perceptual hashing in general and PhotoDNA in particular are industry standards for detecting known CSAM due to their ability to recognize known material that has been slightly altered,[30] identifying PhotoDNA matches requires access to the actual image data. As noted above, images referenced in the LAION datasets frequently disappear, and PhotoDNA was unable to access a high percentage of the URLs provided to it.

To augment this, we used the laion2B-multi-md5, laion2B-en-md5 and laion1B-nolang-md5 datasets[31] datasets. These include MD5[32] cryptographic hashes[33] of the source images, and cross-referenced entries in the dataset with MD5 sets of known CSAM obtained from NCMEC. This approach cannot match images that have been modified in any way since hashing, but can detect some images that

29. Gowthami Somepalli et al., *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*, 2022, 9–10, arXiv: 2212.03860 [cs.LG]; Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, 2023, arXiv: 2301.13188 [cs.CR].
30. Farid, "An Overview of Perceptual Hashing."
31. LAION, *LAION eV*.
32. Ronald L. Rivest, "The MD5 Message-Digest Algorithm," *IETF*, April 1992, https://datatracker.ietf.org/doc/html/rfc1321.
33. As opposed to perceptual hashing which can match subtly different versions of the same image, cryptographic hash algorithms such as MD5 only match against exact copies of the image from which they were calculated. In other terms, using a cryptographic hash instead of a perceptual hash has low sensitivity but high specificity.

have gone offline and thus can no longer be accessed by PhotoDNA. It is also faster and more feasible than performing an API call and perceptual hash evaluation for all 5 billion images.

3.3 Leveraging KNN and ML classification for novel CSAM detection

Due to the potentially large number of neighbors resulting from KNN queries, it was infeasible to submit all URLs to C3P for manual review as potential novel CSAM. To narrow candidates, a new dataset was created with only neighbors that were not PhotoDNA matches. Active URLs were downloaded to a transient RAM disk, and subsequently fed into Thorn's ML classifier before being deleted.

4 Summary of results

After evaluating all images above our chosen safety cutoff in the LAION datasets (32,138,129 items), we found a total of 1,679 PhotoDNA matches. The URLs of these PhotoDNA matches were submitted to NCMEC via the PhotoDNA reporting API and to C3P via the Arachnid API; of the ones that were still live and evaluated by C3P, 746 were classified as instances of CSAM or possible CSAM. Using MD5 hashes, we found 495 matches in all datasets combined. 266 of these had already been discovered by PhotoDNA, leaving 229 unique instances found solely by this method.

Positive matches from all datasets were then compiled into two sets of image embeddings for KNN queries: results that were both PhotoDNA matches and manually confirmed as being CSAM by C3P,[34] and those derived from MD5 matches (many of which were not live and could not be verified).

Based on the neighbors of PhotoDNA- and C3P-validated instances, we computed 1,466,368 neighbors. Of these, 395,496 were unique URLs, and 134,083 had not been scanned already as part of our prior analysis. Evaluating these neighbors via PhotoDNA resulted in an additional 162 hits. Using the CSAM classifier provided by Thorn on the remaining neighbors, 575 results were strongly predicted to be CSAM (99% or higher probability). These were submitted to PhotoDNA for scanning, resulting in 18 matches. These matches were subtracted from the original set, and the remaining URLs submitted to C3P via the Arachnid API for review. C3P confirmed 105 images as being CSAM.

98,918 neighbors were computed for the results detected by the MD5 method, with PhotoDNA detecting 167 of the new neighbors as CSAM. These results were reported and removed from the neighbor set and the remaining files were tested against Thorn's CSAM classifier. URLs of 432 High probability candidates were submitted to the C3P for review, with 78 being classified as CSAM or likely CSAM.

A summary of total findings via PhotoDNA, MD5 match and KNN search is shown in Table 2 on the next page, with findings from the Thorn classifier shown in Table 3.

34. Some instances were difficult for C3P to manually verify, either due to age ambiguity or lack of supplementary information. This does not mean they were not CSAM, but we focused on the most easy to confirm instances to give nearest neighbors searches high accuracy.

Table 2: Instances of CSAM detected by PhotoDNA, MD5 and both techniques combined with KNN. Validated matches are those that were manually confirmed by C3P to be CSAM or likely CSAM. Note that many matches were not able to be conclusively classified, and these uncertain results have been excluded from the count of validated results; this does not mean that they are not CSAM, but that examiners did not have sufficient supportive evidence that may have been available to the organization which submitted the hash. Matches that went offline between the time of initial analysis and submission to C3P were not able to be evaluated, and some duplicates were removed.

	Total	Validated
punsafe >= 0.995 + PhotoDNA	1,679	601
KNN + PhotoDNA	162	49
MD5 match	229	N/A
KNN + MD5 + PhotoDNA	167	78
Total	2,237	825

Table 3: Candidates detected as >99% probability of being CSAM by use of KNN and the Thorn ML classifier. Note that there were no common URLs found between the PhotoDNA and MD5-derived sets.

	Total	Validated
KNN of PhotoDNA + Thorn ML	557	105
KNN of MD5 + Thorn ML	432	78
Total	989	183

A number of notable sites were included in these matches, including the CDNs of Reddit, Twitter, Blogspot and WordPress, as well as those of mainstream adult sites such as XHamster and XVideos; suggestions for mitigating these gaps can be found in Section 5.4. A high percentage of hits were also from sites dedicated to "teen models" or nudism, as well as Japanese "junior idol"[35] content.

35. Wikipedia contributors, *Junior idol*, 2023, accessed December 11, 2023, https://en.wikipedia.org/wiki/Junior_idol.

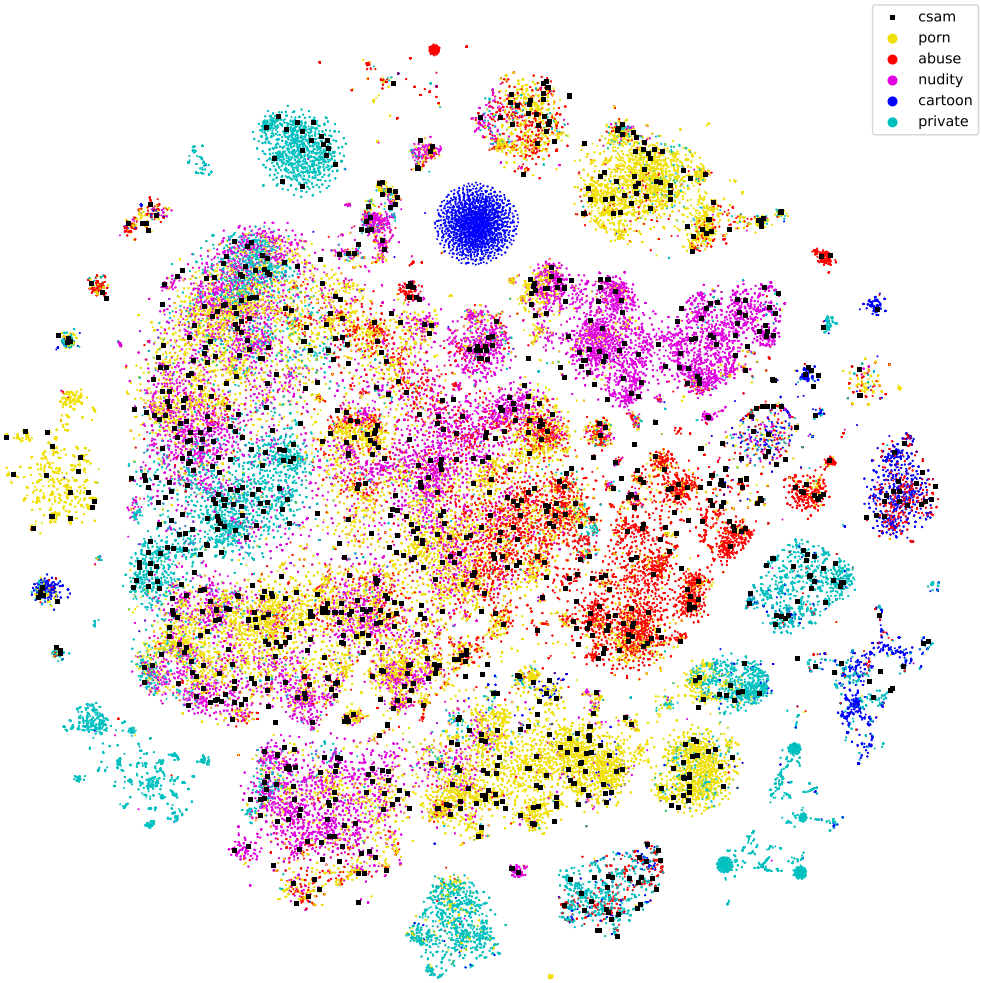

Figure 2: TSNE visualization of 110k images based on their metadata (embeddings), representing 0.002% of the LAION-5B dataset. Confirmed hits are in black along with some of their nearest neighbors from the KNN. Colors are determined from the metadata only, picked using weighted cosine similarity from text embeddings on key categories such as pornography, personal or family photos (privacy), nudity, abuse or cartoon content. Visualization by Alex Champandard.

5 Overall findings and recommendations

We find that having possession of a LAION-5B dataset populated even in late 2023 implies the possession of thousands of illegal images—not including all of the intimate imagery published and gathered non-consensually, the legality of which is more variable by jurisdiction. While the amount of CSAM present does not necessarily indicate that the presence of CSAM drastically influences the output of the model above and beyond the model's ability to combine the concepts of sexual activity and children, it likely does still exert influence. The presence of repeated identical instances of CSAM is also problematic, particularly due to its reinforcement of images of specific victims.

With the knowledge that CSAM is present in several prominent ML training datasets, there are some actionable steps users can take to mitigate the problems posed by the distribution of the content and its inclusion in model training data, as well as ways to prevent such incidents in the future.[36]

5.1 Removing material

There are several aspects to removing the material identified by this project, by order of increasing difficulty:

1. Removal from the original hosting URLs,
2. Removal of the metadata entries in public LAION datasets,
3. Removal of the actual images from the internal LAION reference datasets,
4. Removal of images and references in downloaded copies of the image set in the possession of various researchers, and
5. Removal from the models themselves.

Removal of the source material is already in progress as a result of URLs being identified during this project and reported to NCMEC and C3P. Removal of the entries from the publicly distributed datasets hosted on Hugging Face et al will require coordination with LAION and other parties hosting the material; the same is true for removal from the internal reference sets. Removal from sets in the possession of other, smaller researchers is more difficult: it is not known how many researchers have downloaded the LAION datasets, and even if identified, this would require publishing identifiers which allow people to find illegal materials. Beyond removing the metadata entries, the image embeddings of offending imagery would also need removal; knowledge of these embeddings could also be used by malicious actors to reinforce models intended to produce CSAM.

Removing material from the models themselves is the most difficult task; for images that match known CSAM, the image and text embeddings could be removed from the model, but it is unknown whether this would meaningfully affect the ability of the model to produce CSAM or to replicate the appearance

36. For complementary recommendations for safe development of AI models, see the forthcoming paper by Thorn ("Safety by Design for Generative AI: Combating Child Sexual Abuse").

of specific victims. A more generalizable approach could be that of what has been referred to as concept ablation[37] or concept erasure[38] where the model is trained to partially discard specific concepts. However, unless the concept slated for removal is that of children or nudity as a whole, ablating the concept of CSAM in particular may require access to illegal material itself.

5.2 Implications for training set compilation

The LAION project's use of CLIP filtering for "NSFW" detection is potentially useful, though it is unclear whether this approach outperforms traditional NSFW classifiers.[39] Regardless, it was ultimately decided that only images that were both NSFW and matched "underaged" CLIP filters would be discarded during the crawl process, leaving the dataset with plenty of explicit materials with which trained models could combine the concepts of children and sexual activity. Keywords used in the initial CLIP interrogation were also fairly limited: terms unlikely to yield results such as "underaged" were included in the list along with typos such as "boops" and "breats", while many obvious terms were omitted.[40] Consultation with child safety experts would have helped focus and expand this list. Age estimation models[41] could also potentially assist with the detection of potentially illegal materials.

An obvious gap during the compilation of LAION datasets was that images were not checked against known lists of CSAM. For organizations seeking to compile such datasets, it is advisable to partner with NCMEC, Microsoft, C3P, Thorn or similar sources with access to perceptual or cryptographic hashes of known CSAM. Organizations that have already compiled datasets could also use these same methods and intermediaries to clean their current datasets. Images detected as CSAM should then be submitted to the appropriate entity for triage (NCMEC in the US).

5.3 Alterations to the model training process

Unfortunately, the repercussions of Stable Diffusion 1.5's training process will be with us for some time to come. With the advent of Stable Diffusion XL,[42] there appears to be some degree of increasing user interest[43] in this presumably

37. Nupur Kumari et al., *Ablating Concepts in Text-to-Image Diffusion Models*, 2023, arXiv: 2303.13516 [cs.CV].
38. Rohit Gandikota et al., *Erasing Concepts from Diffusion Models*, 2023, arXiv: 2303.07345 [cs.CV]; Rohit Gandikota et al., *Unified Concept Editing in Diffusion Models*, 2023, arXiv: 2308.14761 [cs.CV].
39. For example, Laborde (*Deep NN for NSFW Detection*) and Bumble Inc. (*Private Detector*).
40. Schuhmann, *crawlingathome*.
41. See, for example, Serengil and Ozpinar ("HyperExtended LightFace: A Facial Attribute Analysis Framework").
42. Dustin Podell et al., *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*, 2023, arXiv: 2307.01952 [cs.CV].
43. This is based upon increasing prevalence of SDXL checkpoints on CivitAI as well as user preference comparisons from Podell et al. (*SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*, 2). However, note that the user preference comparisons were made based on the official base Stable Diffusion models, as opposed to some of the more "advanced" community models based on Stable Diffusion 1.5.

"safer"[44] model.

We have previously proposed[45] that models trained on erotic content not be trained on material depicting children; this limits the ability of models to conflate the two types of material. Indeed, given the regulatory scrutiny regarding gathering data on children (e.g. COPPA[46]), images of children should arguably be excluded from generalized training sets entirely.

5.4 Content hosting platforms

Presumably, several of the platforms hosting content identified as being CSAM do implement some detection methodologies. However, due to the asynchrony between the time content is uploaded, the time a certain piece of content is added to industry hash databases and the time at which a given service provider commenced proactively scanning for CSAM, there is likely a need for platform providers to re-scan material they host.[47] For example, if CSAM content was uploaded to a platform in 2018 and then added to a hash database in 2020, under commonly used methods that content could potentially stay on the platform undetected for an extended period of time.

There are several methods that might be implemented to address this issue:

1. When implementing a CSAM detection technology, ensure that it is retroactively applied to all content currently on the platform. Remember to include thumbnail/preview images as well as full-scale.
2. Store PDQ[48] hashes of all content on the platform (including retroactively). Periodically re-scan PDQ hashes against industry PDQ hash databases to detect material added to hash databases after the content was uploaded.

5.5 Model and dataset hosting platforms

For platforms such as Hugging Face which distribute the training data for models, a more streamlined mechanism for reporting and removing links to CSAM should be implemented. Platforms that host and distribute models trained on LAION-5B or derivative datasets (e.g. CivitAI) should also implement a mechanism for reporting and removing models and augmentations known to produce or substantially assist the creation of CSAM.

44. Given past challenges with explicit content, this is is only presumed; there is no publicly available information about exactly what images were used to train SDXL and what safety measures were in place, and its model card is silent on this issue.
45. Thiel, Stroebel, and Portnoff, "Generative ML and CSAM: Implications and Mitigations."
46. Federal Trade Commission, *Children's Online Privacy Protection Rule ("COPPA")*, accessed November 27, 2023, https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.
47. This problem will increase in severity given the high volume of computer generated CSAM content; much, if not most, of this content is novel and posted on platforms significantly before it can be added to a hash database.
48. Antigone Davis and Guy Rosen, *Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer*, August 1, 2019, https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

6 Safety and ethical considerations

The techniques used in this project relied primarily on submission of URLs to the PhotoDNA API and C3P Arachnid APIs, as well as comparison of pre-calculated hash values. No known CSAM was downloaded or stored locally as part of this process. Candidates detected by KNN searches which were not PhotoDNA or MD5 matches were downloaded to a RAM disk on a transient VM and immediately deleted after classification, with the URLs of any matches submitted to C3P.

All URLs detected to contain CSAM were reported to both NCMEC and the C3P to liase with law enforcement and service providers. Records of the URLs were kept in a secure storage environment separate from SIO's main cloud infrastructure.

7 Conclusion

The material detected during this process is inherently a significant undercount due to the incompleteness of industry hash sets, attrition of live hosted content, lack of access to the original LAION reference image sets, and the limited accuracy of "unsafe" content classifiers. We have, however, attempted to be as thorough as possible given the limitations posed by the massive dataset size, and hope that these techniques may be of some use in addressing similar problems in other datasets (or indeed, in the developing of improved content classifiers).

Web-scale datasets are highly problematic for a number of reasons even with attempts at safety filtering. Apart from CSAM, the presence of non-consensual intimate imagery (NCII)[49] or "borderline" content[50] in such datasets is essentially certain—to say nothing of potential copyright and privacy concerns. Ideally, such datasets should be restricted to research settings only, with more curated and well-sourced datasets used for publicly distributed models.

We are now faced with the task of trying to mitigate the material present in extant training sets and tainted models; a problem gaining urgency with the surge in the use of these models to produce not just generated CSAM,[51] but also CSAM and NCII of real children,[52] often for commercial purposes.[53] The most obvious solution is for the bulk of those in possession of LAION-5B-derived training sets to delete them or work with intermediaries to clean the material. Models based on Stable Diffusion 1.5 that have not had safety measures applied to them should be deprecated and distribution ceased where feasible.

49. INHOPE, "What is NCII?," February 17, 2023, https://inhope.org/EN/articles/what-is-ncii.

50. Usually used to mean content which portrays clothed children but in a sexualized or easily sexualizable manner.

51. Thiel, Stroebel, and Portnoff, "Generative ML and CSAM: Implications and Mitigations."

52. Guy Hedgecoe, "AI-generated naked child images shock Spanish town of Almendralejo," *BBC News*, September 24, 2023, https://www.bbc.com/news/world-europe-66877718; Tim McNicholas, "New Jersey high school students accused of making AI-generated pornographic images of classmates," *CBS News*, Updated on November 2, 2023, https://www.cbsnews.com/newyork/news/westfield-high-school-ai-pornographic-images-students/.

53. Kolina Koltai, "AnyDream: Secretive AI Platform Broke Stripe Rules to Rake in Money from Nonconsensual Pornographic Deepfakes," *Bellingcat*, November 27, 2023, https://www.bellingcat.com/news/2023/11/27/anydream-secretive-ai-platform-broke-stripe-rules-to-rake-in-money-from-nonconsensual-pornographic-deepfakes/.

Acknowledgements

We would like to thank the following people and organizations for their help with this project:

- Alex Champandard for technical assistance and initial project inspiration.
- The Canadian Centre for Child Protection (C3P) for validating our results and initiating takedown requests.
- Microsoft for high-volume access to the PhotoDNA API.
- Thorn for access to their classifier to identify candidates for evaluation.
- The National Center for Missing and Exploited Children (NCMEC) for access to the MD5 hash databases used to detect missing content and for triage of detected instances.

References

Beaumont, Romain. *Clip Retrieval: Easily compute clip embeddings and build a clip retrieval system with them*. GitHub repository, 2022. https://github.com/rom1504/clip-retrieval.

———. "Image embeddings," July 20, 2020. https://rom1504.medium.com/image-embeddings-ed1b194d113e.

———. "Large Scale OpenCLIP: L/14, H/14 and G/14 Trained on LAION-2B," September 15, 2022. https://laion.ai/blog/large-openclip.

Bumble Inc. *Private Detector*. GitHub repository. Accessed October 3, 2023. https://github.com/bumble-tech/private-detector.

Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. *Extracting Training Data from Diffusion Models*, 2023. arXiv: 2301.13188 [cs.CR].

Carlsson, Fredrik, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. "Cross-lingual and Multilingual CLIP." In *Proceedings of the Language Resources and Evaluation Conference*, 6848–54. Marseille, France: European Language Resources Association, June 2022. https://aclanthology.org/2022.lrec-1.739.

Common Crawl - Open Repository of Web Crawl Data. Accessed September 15, 2023. https://commoncrawl.org.

Davis, Antigone, and Guy Rosen. *Open-Sourcing Photo- and Video-Matching Technology to Make the Internet Safer*, August 1, 2019. https://about.fb.com/news/2019/08/open-source-photo-video-matching/.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. 2009. https://doi.org/10.1109/CVPR.2009.5206848.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, et al. *Beyond English-Centric Multilingual Machine Translation*, 2020. arXiv: 2010.11125 [cs.CL].

Farid, Hany. "An Overview of Perceptual Hashing." *Journal of Online Trust and Safety* 1, no. 1 (October 2021). https://doi.org/10.54501/jots.v1i1.24.

Federal Trade Commission. *Children's Online Privacy Protection Rule ("COPPA")*. Accessed November 27, 2023. https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

Gandikota, Rohit, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. *Erasing Concepts from Diffusion Models*, 2023. arXiv: 2303.07345 [cs.CV].

Gandikota, Rohit, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. *Unified Concept Editing in Diffusion Models*, 2023. arXiv: 2308.14761 [cs.CV].

Hedgecoe, Guy. "AI-generated naked child images shock Spanish town of Almendralejo." *BBC News*, September 24, 2023. https://www.bbc.com/news/world-europe-66877718.

INHOPE. "What is NCII?," February 17, 2023. https://inhope.org/EN/articles/what-is-ncii.

Koltai, Kolina. "AnyDream: Secretive AI Platform Broke Stripe Rules to Rake in Money from Nonconsensual Pornographic Deepfakes." *Bellingcat*, November 27, 2023. https://www.bellingcat.com/news/2023/11/27/anydream-secretive-ai-platform-broke-stripe-rules-to-rake-in-money-from-nonconsensual-pornographic-deepfakes/.

Kumari, Nupur, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. *Ablating Concepts in Text-to-Image Diffusion Models*, 2023. arXiv: 2303.13516 [cs.CV].

Laborde, Gant. *Deep NN for NSFW Detection*. GitHub Repository. Accessed October 3, 2023. https://github.com/GantMan/nsfw_model.

LAION. *LAION eV*. Hugging Face repository. Accessed October 2, 2023. https://huggingface.co/laion.

McNicholas, Tim. "New Jersey high school students accused of making AI-generated pornographic images of classmates." *CBS News*. Updated on November 2, 2023. https://www.cbsnews.com/newyork/news/westfield-high-school-ai-pornographic-images-students/.

Nezhurina, Marianna, Romain Beaumont, Richard Vencu, and Christoph Schuhmann. *LAION Translated: 3B Captions Translated to English from LAION5B*, September 15, 2022. https://laion.ai/blog/laion-translated/.

Podell, Dustin, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*, 2023. arXiv: 2307.01952 [cs.CV].

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. *Learning Transferable Visual Models From Natural Language Supervision*, 2021. arXiv: 2103.00020 [cs.CV].

Rivest, Ronald L. "The MD5 Message-Digest Algorithm." *IETF*, April 1992. https://datatracker.ietf.org/doc/html/rfc1321.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*, 2022. arXiv: 2205.11487 [cs.CV].

Schuhmann, Christoph. *CLIP-based-NSFW-Detector*. GitHub repository. Accessed September 15, 2023. https://github.com/LAION-AI/CLIP-based-NSFW-Detector.

———. *crawlingathome*. GitHub repository. Accessed September 15, 2023. https://github.com/christophschuhmann/crawlingathome-worker/blob/master/crawlingathome.py#L475.

Schuhmann, Christoph, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, et al. "LAION-5B: An open large-scale dataset for training next generation image-text models." In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2022. https://openreview.net/forum?id=M3Y74vmsMcY.

Schuhmann, Christoph, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. *LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs*, 2021. arXiv: 2111.02114 [cs.CV].

Serengil, Sefik Ilkin, and Alper Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework." In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, 1–4. IEEE, 2021. https://doi.org/10.1109/ICEET53442.2021.9659697.

Somepalli, Gowthami, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*, 2022. arXiv: 2212.03860 [cs.LG].

Stability AI. *Stable Diffusion v2 Model Card*, November 23, 2022. https://huggingface.co/stabilityai/stable-diffusion-2.

———. *Stable Diffusion v2-1 Model Card*, December 7, 2022. https://huggingface.co/stabilityai/stable-diffusion-2-1.

Thiel, David, Melissa Stroebel, and Rebecca Portnoff. "Generative ML and CSAM: Implications and Mitigations." *Stanford Digital Repository*. https://doi.org/10.25740/jv206yg3793.

Thorn. "Safety by Design for Generative AI: Combating Child Sexual Abuse." Forthcoming, January 2024.

Vincent, James. "Stable Diffusion made copying artists and generating porn harder and users are mad." *The Verge*, November 24, 2022. https://www.theverge.com/2022/11/24/23476622/ai-image-generator-stable-diffusion-version-2-nsfw-artists-data-changes.

Wikipedia contributors. *Junior idol*, 2023. Accessed December 11, 2023. https://en.wikipedia.org/wiki/Junior_idol.

Stanford | Internet Observatory
Cyber Policy Center

The Stanford Internet Observatory is a cross-disciplinary program of research, teaching and policy engagement for the study of abuse in current information technologies, with a focus on social media. The Stanford Internet Observatory was founded in 2019 to research the misuse of the internet to cause harm, formulate technical and policy responses, and teach the next generation how to avoid the mistakes of the past.

Stanford | Internet Observatory
Cyber Policy Center