Generative ML and CSAM: Implications and Mitigations

David Thiel, Melissa Stroebel and Rebecca Portnoff

June 24, 2023

THORN
Stanford | Internet Observatory
Cyber Policy Center

Contents

1 Introduction

The adversarial use of generative machine learning models has been recognized in the study of mis- and disinformation for quite some time—historically, Generative Adversarial Networks (GANs) have been used to generate realistic-looking (although relatively easily detectable) avatars for fake accounts on social media services. With the release of freely available conditional Diffusion Models[1] (DMs) such as DALL-E,[2] Midjourney[3] and Stable Diffusion,[4] visual generative machine learning models became more flexible and user-friendly, generating elaborate scenes based on user-supplied textual prompts. Throughout much of 2022, these generative ML models were often glitch-prone and somewhat difficult to control—potentially dismissable as toys for hobbyists or limited to imitative artwork.

Initial diffusion models were released with fairly limited safety controls. While some controls have been added after the fact in ad-hoc fashion to prevent political misinformation[5] or adult content,[6] prior versions of models like Stable Diffusion that were partially trained on adult content remain in heavy use. A community has developed to attempt to advance its ability to generate flexible and realistic new adult content.

In just the first few months of 2023, a number of advancements have greatly increased end-user control over image results and their resultant realism, to the point that some images are only distinguishable from reality if the viewer is very familiar with photography, lighting and the characteristics of diffusion model outputs. Near-realistic adult content is currently distributed online in public and private web and chat forums. This advancement has also enabled another type of imagery: material in the style of child sexual exploitation content.

With the pace of diffusion model development, it is likely that in under a year it will become significantly easier to generate adult images that are indistinguishable from actual images.[7] This presents a number of social challenges, not the least of which is Child Sexual Abuse Material (CSAM) that cannot be definitively distinguished as being photographic or computer-generated.

In a scenario where highly realistic computer-generated CSAM (CG-CSAM) becomes highly prevalent online, the ability for NGOs and law enforcement to investigate and prosecute CSAM cases may be severely hindered. Currently, the prevalence of CG-CSAM is small but growing. Based on an internal study by Thorn, less than 1% of CSAM files shared in a sample of communities dedicated to child sexual abuse are photorealistic CG-CSAM, but this has increased consistently

[1] Yang et al. 2023.

[2] https://openai.com/research/dall-e

[3] https://www.midjourney.com

[4] https://github.com/CompVis/stable-diffusion

[5] Stanley-Becker and Harwell 2023.

[6] Vincent 2022.

[7] This inability to distinguish is exacerbated by adult images that are already post-processed to some degree using visual filters, airbrushing, or even enhancement with diffusion models themselves.

since August 2022. When focusing on only CG-CSAM in these communities, Thorn finds that approximately 66% are highly photorealistic, but can currently be visually distinguished as being generated (e.g. due to roughness of skin edges, or highly pixelated areas).

In this report, we examine the current state of the art of visual generative models, the hurdles currently preventing widespread proliferation, and techniques that may soon overcome those hurdles. We also look at the societal implications and technical and policy countermeasures to mitigate the problem. Our primary focus is on Stable Diffusion due to its open-source nature and active community, but recommendations apply to all similar models.

2 What is a diffusion model?

Essentially, a diffusion model is a form of machine learning trained by images with progressive layers of noise added (forward diffusion), until they become a visual representation of Gaussian noise. After training, the model is able to reverse the process, progressively de-noising until producing an image (reverse diffusion). Commonly used diffusion models are specifically “conditional” diffusion models, which are trained with image data tagged with textual detail about the contents of each image. These models subsequently use text prompts or a source image to help guide the model from a random seed to an image with characteristics specified by the input.

Models can also be conditioned with negative prompts, which specify what types of effects are undesirable in the output—this can be to exclude a certain art style (e.g. “cartoon”, “monochrome”), image artifacts (e.g. “extra fingers”, “distorted”), or image components (e.g. “clothed”, “outdoors”).

2.1 Historical limitations of diffusion models

There are a number of traits of current diffusion models that often allow for discerning them from real-world imagery:

- Lack of skin imperfections or texture
- Inaccurate hands or joint poses
- Shadows that do not reflect real-world lighting conditions
- Difficulty depicting wet surfaces or moisture
- Jumbled logos or signs
- Asymmetry of ears or eyes
- Turnaround time for image generation

Depending on desired output, some of these may not be a problem—much current computer-generated sexual imagery is intended to be “hyperrealistic” or resembling cartoons (for example, in an anime style). However, there are communities attempting to develop easily-accessible fully realistic adult content models and augmentations that can overcome these limitations. With these

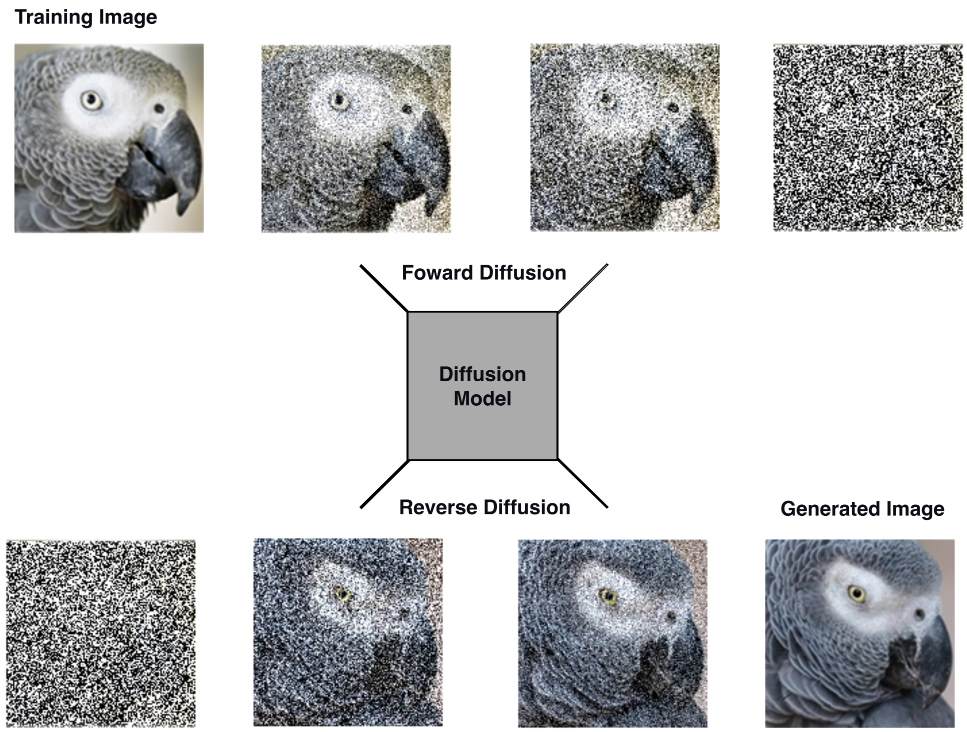

Figure 1: A simplified representation of the diffusion model training and generation process. Collections of images are gradually noised during the training process to produce the diffusion model. The resultant generative model then starts from pure noise, de-noising to produce a visual representation.

augmentations, many prompts that produce sexually explicit outputs can be easily changed to output children—in fact, multiple models focused on adult content suggest putting "child" in the negative prompt to prevent accidental CG-CSAM.

Attaining full realism also currently takes a significant amount of trial and error, with image generation requiring anywhere from 30 seconds to 10 minutes on consumer-grade CPU hardware to several seconds on a high-end GPU (although significantly longer if upscaling is performed or high batch sizes are used). As both hardware and software technology advances, these limitations will only lessen, potentially allowing for near-realtime adjustments.

3 Advances in the diffusion model ecosystem

The first available methods for fine-tuning Stable Diffusion models were resource-intensive and required high-powered GPUs to train. For many users, this necessitated use of third-party environments such as Google Colab[8] to perform

[8] https://colab.research.google.com

training, with potential oversight making it less suitable for generating adult content. Newer techniques require far fewer resources, putting training within reach of even casual hobbyists.

Using any of these techniques can help produce realistic-looking sexually explicit imagery when used in conjunction with the standard Stable Diffusion 1.5 model. They can also be trained on existing CSAM material to the point that the output can resemble the actual victims it was trained upon—a result which members of online CSA communities purport to have achieved. We present a non-comprehensive list of such techniques below.

3.1 DreamBooth

The primary tradeoff when choosing a technique to fine-tune an existing Stable Diffusion model is flexibility versus accuracy. If the ability to output a wide range of styles or scenes is a goal, DreamBooth[9] is the preferred method. With this technique, a small set of training images of a subject are used to create a new model that can recontextualize the subject or expand its capabilities to a new style. For example, a few images of a corgi puppy enables the generation of images featuring a corgi puppy in distinct scenarios or places. The resulting "checkpoint" file is a point in time snapshot of the entire model, with DreamBooth augmentations.

The primary drawback of DreamBooth as a technique is the amount of video RAM and disk space required, as well as the size of the output model (typically 2–4G). While the resources required for earlier implementations of DreamBooth required fairly expensive hardware, current implementations can be run with a video card under 16G, obtainable at a price of under $1000.

3.2 Textual inversion and LoRA

Textual inversion and LoRAs are two techniques that can be used for similar purposes, differing in the underlying technique, required resources and file size of the model augmentation they produce.

With textual inversion,[10] samples of imagery can be combined with descriptions to create new, machine-readable "words" that can be added to a positive or negative textual prompt. This can be used to steer output toward a certain artistic style, character, or body type. When used as a negative prompt, textual inversion can help correct for some of the visual defects common with diffusion models, such as malformed hands or eyes.[11]

Low-Rank Adaptations[12], or LoRAs, provide a lightweight way to train add-on model weights that steer output in a particular direction. In adult content creation, this can include touch-ups for specific parts of anatomy or suggesting particular

[9] Ruiz et al. 2022.

[10] Gal et al. 2022.

[11] See, for example, https://huggingface.co/datasets/Nerfgun3/bad_prompt.

[12] Hu et al. 2021.

poses and acts. A wide array of community-produced LoRAs[13] have proliferated in the early months of 2023, along with user-supplied prompts and seeds to replicate and customize results. LoRAs are also referenced in communities dedicated to child sexual abuse for their ease of use to fine tune Stable Diffusion models with existing content of victims, to generate more content of those victims.

LoRAs can also be used with “inpainting”, which applies a mask to a certain part of an image and uses textual prompts to do visual revisions or touch-ups. In CSAM communities, Thorn has observed bad actors discussing the use of inpainting to incorporate sexualized facial expressions on the child’s face.

LoRAs and textual inversions can also be used to replicate the likeness of a particular person or fictional character, giving results that may be difficult to achieve with a base model (particularly if the person is not well-known). This has been used to create realistic images of living people or children, and can also be used to adjust the character’s age (see Figure 2 for a demonstration of likeness-matching and age manipulation).

In contrast to DreamBooth checkpoints, both TIs and LoRAs are small enough to be shared in a group chat. Multiple TIs and/or LoRAs can be used simultaneously with different weights to adjust output.

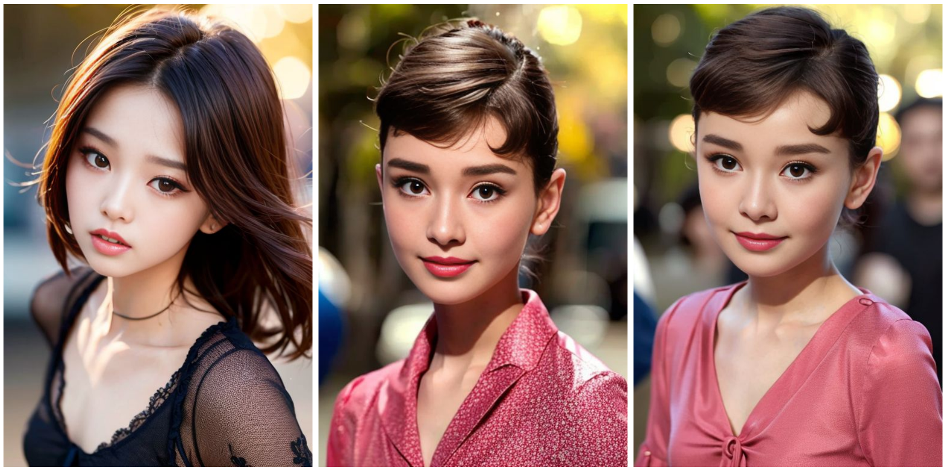

Figure 2: Left: A woman generated with a popular Stable Diffusion model. Center: The same prompt, but with a LoRA to make the output moderately resemble Audrey Hepburn. Right: Addition of a textual inversion to make the resulting character appear younger.

3.3 ControlNet and OpenPose

ControlNet[14] is a model which allows for fine-tuning and restyling of images via text prompts and visual inputs. For example, it can extract visual or human pose data[15] from existing photos. Users can then use text prompts to generate new

[13] https://civitai.com

[14] Zhang and Agrawala 2023.

[15] Cao et al. 2019.

images matching the original pose (see Figure 3).

This allows for fine control over the activities of subjects, as well as restyling them aesthetically—and as time passes, adjusting poses and other visual aspects will get closer and closer to being real-time. Pose files are currently distributed in the open source generative ML community, which offers a potentially more lightweight way to achieve some results previously requiring the use of a LoRA.

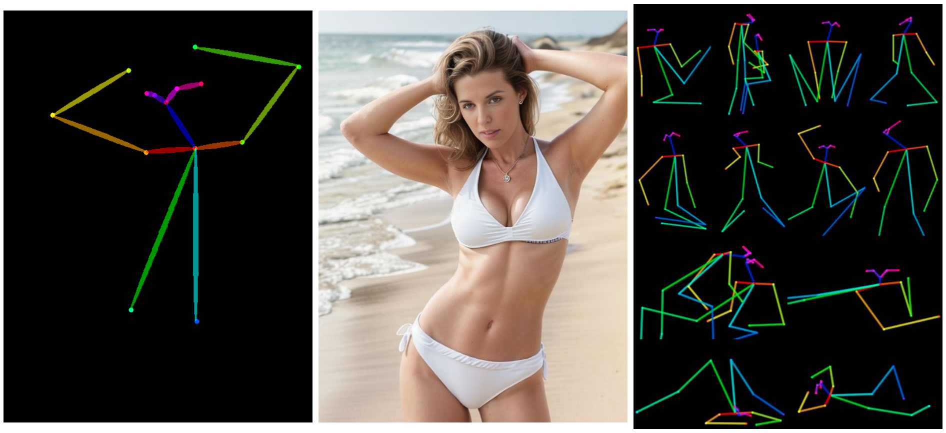

Figure 3: Left: An OpenPose "skeleton" pose. Center: A scene using that pose, generated by Stable Diffusion in conjunction with ControlNet. Right: A variety of different OpenPose poses, some of which are potentially usable for creating explicit content. OpenPose skeletons from CivitAI.

4 Consequences of realistic CG-CSAM proliferation

There are legitimate uses for all of these tools. However, they are already being used to generate realistic CG-CSAM, so understanding the implications is important for policymakers, platforms, and the public alike.

There are a variety of actual and potential societal and technical outcomes resulting from the widespread ability to produce CG-CSAM. For completeness, it is worth noting some have argued, under the right controls, such material could be used to reduce offender risk in some instances. Some theorize that the use of CG-CSAM in place of CSAM produced from the non-virtual abuse of living children could serve a preventative purpose—potentially for treatment/impulse management of those identifying with a sexual attraction to minors. However, neither the viability nor efficacy of such a practice has been sufficiently studied and many warn that, for some, this material could have an adverse effect—lowering barriers of inhibition or contributing to existing fantasies of real-world abuse.[16]

Several decidedly negative outcomes have been observed and pose a high risk to child safety as the availability of CG-CSAM grows. One likely scenario is that the advent of realistic CG-CSAM generates hundreds of thousands of reports

[16] Christensen, Moritz, and Pearson 2021.

to technology platforms, NGOs handling CSAM cases, and law enforcement, thus overloading the ability of companies and organizations to effectively handle reporting and investigations. The investigators will have the added challenge of determining whether the victim in the scenario is in fact a real person.

Another active scenario is the further re-victimization of children depicted in CG-CSAM, as textual prompts and fine-tuned models enable the generation of more CG-CSAM matching the likeness of the child in the original material. This CG-CSAM uses the original abuse material to produce content with new poses and sexual acts, including egregious content like sexual violence.

Finally, cases are already being reported where AI generative technologies are being employed to facilitate the grooming and sextortion of minor victims. In June 2023, the FBI issued a warning[17] about this threat, warning the public the technology may be used to generate explicit images of minors from benign images located online. In addition to creating explicit imagery to extort new victims who have not shared sensitive imagery, this technology risks being utilized to scale existing sextortion schemes, producing imagery and targeting potential victims at a previously unseen rate.

4.1 Legal concerns

18 U.S. Code § 2252A[18] uses a standard[19] prohibiting any visual depiction of CSA that is “virtually indistinguishable” from a minor engaging in sexual conduct. The definition it uses specifically references computer-generated material as being in scope. Additionally, 18 U.S. Code § 1466A[20] states that any depiction of a minor that both contains sexually explicit conduct and is obscene can be prosecuted with the same penalties as 2252A.

As such, the current status of CG-CSAM appears to be that prosecution under 1466A is possible for any representation deemed both graphic and obscene,[21] and under 2252A if that material is indistinguishable. Now that CG-CSAM has reached the point that it may be truly virtually indistinguishable from photographic CSAM and the methods to generate material are widely available, the application of some of these laws may be tested.

Realistic CG-CSAM also presents obvious difficulties when it comes to prosecutions for CSAM possession; while the tactic of a defendant insisting that real CSAM is artificial has long been anticipated, in general, the appearance of a child being abused has been sufficient for prosecution. In a world where such a tactic could become commonplace, alternatives may need to be tested that do not require positive identification of a real-world victim (or worse, that would require such a victim or their family to testify).

[17] FBI 2023.

[18] 18 U.S. Code § 2252A 2018.

[19] 18 U.S. Code § 2256 2018.

[20] 18 U.S. Code § 1466A 2003.

[21] See *U.S. v. Whorley* (2005).

The situation is further complicated by the fact that CG-CSAM can easily resemble a real-world victim of CSA, as well as non-consensually borrow the likeness of any other child—both of which can inflict harm on that person (and both of which are specifically prohibited by 18 USC § 2256(8)(C)).

Beyond the potential problem of insisting that real-world CSAM is a fully computer-generated scene, some diffusion model techniques could allow modification of scenes to help obscure evidence or confound investigations. For example, the producer or a person in possession of a picture depicting an actual victim might modify it to have some of the quirks of a diffusion model (e.g., those listed in Section 2.1 on page 3). While this might not be fully ultimately effective, it could impede investigations.

4.2 Moderation concerns

Most major tech platforms will proactively ban anything resembling CSAM, but those that do not—as well as platforms that are decentralized—present problems for service administrators, moderators and law enforcement. Depending on the location of a server, this may or may not put the operator at legal risk—it may also be legally unclear to operators what law even applies, and whether that is dependent on realism.

The potential increase in volume of CSAM could have further negative effects for those tasked with moderating content on platforms—moderators are already exposed to a significant amount of problematic material and are expected to make very fast judgement on how to action said material. An increased load of potentially traumatic material that is difficult to distinguish as computer generated can make moderation work more stressful and more damaging to mental health.

5 Near-term mitigations

When exploring mitigations, it is important to keep in mind a broad range of potential bad actors, ranging from the technical expert, to the basic user. While there will always be bad actors who can circumvent various mitigations, partial solutions that capture “low-hanging fruit” may still be worth pursuing. Mitigations can include alterations to the way models are built and implemented to prevent them from producing CG-CSAM, as well as mechanisms to identify CG-CSAM and potentially reduce the burden on platforms, NGOs and investigators.

In a similar vein, platforms that host models, e.g. HuggingFace or CivitAI, can have community standards regarding how much flexibility hosted models have. These platforms host some models, LoRAs and other augmentations that are made explicitly for adult content. Platforms could require that models geared toward erotic content are similarly weighted against producing representations of children before consenting to host and distribute these tools. Data collected by Thorn consistently indicates that of the specific models named in communities dedicated to child sexual abuse, the overwhelming majority are Stable Diffusion-based models fine-tuned for general NSFW use, predominantly obtained via CivitAI.

5.1 Biasing models against child nudity

In a sense, bias is at the core of all generative machine learning: their function is to reproduce content based on statistical biases in their training data. This lends itself to some well-documented problems, such as representing doctors as always being men, or women as frequently naked.[22]

Just as weights can be used with a pre-built model after the fact—for example, using embeddings in negative prompts to reduce undesirable visual artifacts—the weights of originally trained models themselves could be biased heavily against outputting child nudity or any sexual content involving children. Just as with later models released by Stable Diffusion, training material may exclude pornographic content, making it more difficult to use the models for sexual purposes. In similar fashion, training material can be cleaned of CSAM, using both hashlists of known abuse content and ML/AI solutions. Red teaming exercises can also help indicate edge cases and weak points of the model, to inform further iteration and refinement.

5.2 Watermarking and content provenance

Stable Diffusion by default uses an invisible watermark[23] embedded in the graphical content image files, which makes it difficult to strip out and allows identifying content produced by the model in a manner partially resistant to image reprocessing, resizing or cropping. This is a commendable measure that is notably absent from commercial diffusion models such as DALL-E and Midjourney, but because this process is performed as a separate step after image generation, it is not particularly robust. Given the open-source nature of Stable Diffusion software, the code generating watermark can easily be removed by the end-user. In fact, the most popular version of software used to run Stable Diffusion locally has removed watermarking entirely.[24]

Even if this watermarking technique were consistently applied, there exists the possibility that real images could have the Stable Diffusion watermark applied to them, fooling detection mechanisms and causing different types of confusion. This makes currently implemented watermarking useful for some applications (for example, excluding DM-generated content from training data),[25] but of limited utility when it comes to differentiating CG-CSAM.

[22] This bias is so strong that even when “nude” is provided as a negative prompt, some models will still frequently output nudity.

[23] Zhang et al. 2019.

[24] https://github.com/AUTOMATIC1111/stable-diffusion-webui/issues/2803

[25] Some artists have used the Stable Diffusion watermark on their own work deliberately, in the hope that it may exclude their content from being a training input for future models; see https://github.com/eballai/NoAI.

Newer methods such as Stable Signature[26] can train a watermark into the decoder of a model itself, making it more difficult to remove as well as difficult to imitate; the entity that trains the model would then retain a secret copy of a trained watermark extractor (see illustration in Figure 4). This could ensure that even open-sourced models come with watermarking by default, as opposed to as a voluntary post-processing step. These methods can also identify different copies of a model, potentially identifying the user or organization that generated the image. Both the detection mechanism (i.e., whether the image was generated by the DM) and the identification mechanism (which copy of the model generated it) return probabilistic results and can withstand some degree of image modification.[27]

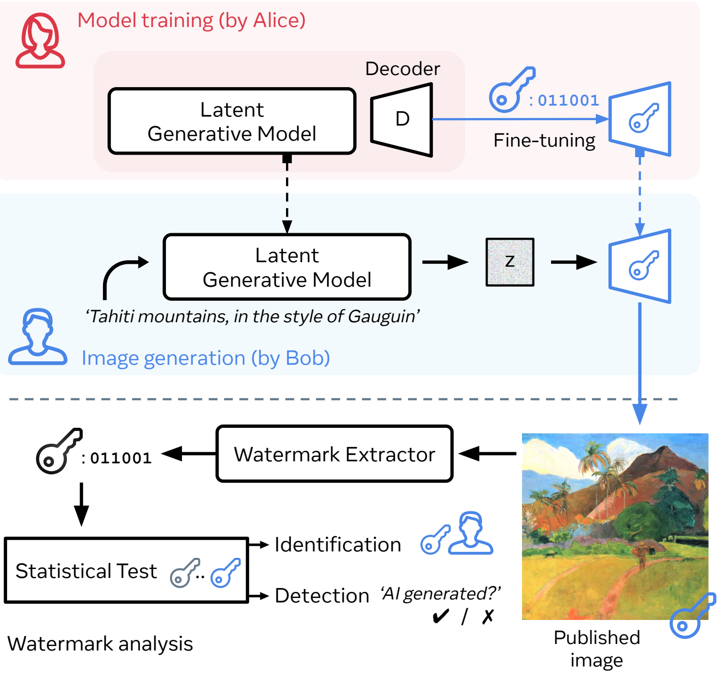

Figure 4: The watermark training and extraction process of Stable Signature. Figure from Fernandez et al. (2023, 1), used with permission.

Lastly, there is another potential mechanism for verifying that content was generated: the exact prompt, model, seed, LoRAs and other data are often embedded in the "User Comment" EXIF field in JPEGs or the "Parameters" field in PNG files. While this obviously can be forged (and may be stripped out during later processing), it should theoretically allow a third party to use the same models and parameters to generate the exact same image if the model and extensions are available to the investigator. This technique is obviously quite difficult to scale, but may be of use in some high-profile or more difficult cases.

[26] Fernandez et al. 2023.

[27] Note that both of these tasks become more difficult in the case of an inpainted image.

5.3 Passive detection mechanisms

Detecting imagery generated by DMs is often discussed in a context of detecting content fabricated by a geopolitical adversary, distributing propaganda to affect political or social outcomes. As such, these adversaries may have a vested interest in generating content that cannot be differentiated from reality,[28] and may invest resources in evading detection. This results in a race where image classifiers are used to train image generators to evade detection.

In the case of CG-CSAM, this adversarial position is unlikely to be present—in most cases, the producer only has incentive to make imagery visually indistinguishable from reality, rather than legally or technically indistinguishable. This makes the possibility of detection more tenable. Even so, if new models and tools are released that continue to make the production of CG-CSAM easier, adversaries may pivot to using those new models. If existing detection solutions are not robust to changing model architectures, new detection solutions may need to be developed to still accurately distinguish content.

Much previous work to detect artificially generated imagery has focused on detecting images generated by GANs. While feasible, these methods are not widely generalizable[29] to images produced by DMs[30] and often only detect the use of one particular generation model out of many possibilities. As such, while the accuracy of detection mechanisms trained on a specific model or generation algorithm can be fairly accurate, a comprehensive detection system would need to pass an image through multiple detection passes for every known model, making it computationally expensive and potentially prohibitive to perform at scale. A possible compromise could be performing a detection pass only when content is flagged as potentially problematic, either by a user report or by a separate ML classifier (such as nudity detection).

Newer methods to detect DM-generated imagery from multiple models at once[31] appear promising; multi-level approaches have also been developed which can distinguish GANs from DMs as well as detect the generative model used.[32] There are also additional techniques that have been developed[33] to detect the use of inpainting.[34] While detection rates of these techniques may change given changes to weights and visual improvements from LoRAs, it is an important area for future research and investment.

[28] Though they may have an incentive to make real material look fake, as discussed in Section 4.1.

[29] Corvi et al. 2022.

[30] Although interestingly, methods trained on DMs appear to be more generalizable to GANs; see Ricker et al. (2023).

[31] Wang et al. 2023.

[32] Guarnera, Giudice, and Battiato 2023.

[33] Wu and Zhou 2022.

[34] Albeit with corresponding effort to evade this detection; see Dou, Feng, and Qian (2023).

5.4 Active monitoring of CG-CSAM production networks

Just as CSAM production and distribution networks are actively monitored by law enforcement, CG-CSAM production forums could be monitored and the perceptual hashes[35] of the material added to separate hash sets. This could allow platforms to detect and remove future uploaded CG-CSAM content; platforms themselves could also contribute to these hash sets, as is currently the case with other hash sharing systems like GIFCT.[36] This material can also be analyzed for trends in models and parameters, as well as potentially used for training of detection models.

Note that while some CG-CSAM which does not qualify as photorealistic may be legally classified as obscenity and thus not subject to the same legal reporting standards as CSAM under 2256, instances of photorealistic CG-CSAM that could reasonably be a depiction of a known victim will need to be reported by platforms. We propose that this warrants the expansion of industry classification and categorization systems.

5.5 Changes to industry CSAM classifications

Globally, multiple classification systems[37] are used by the child safety community and service providers to indicate the severity of content, based on the age of the subject(s) and the acts involved. This could be expanded to include additional criteria for determining severity of CG-CSAM. One possibility would be adding a “C“ classification system to characterize whether content is:

- Computer generated
- Indistinguishable from photo representations
- Portrays explicit sexual activity
- Is modeled after an extant person or known victim

As with existing classifications, an indicator of the estimated age group of the subject could also be included.

5.6 Technical collaboration

While competition is fierce in the field of generative ML, many of these measures benefit from active collaboration between the largest players in the space; for example, each vendor implementing totally different watermarking mechanisms increases the computational cost of automated detection systems, such that the possible existence of a watermark would need to be calculated individually for each watermark implementation.

Collaboration between platform providers is also historically limited, but could be siginificantly beneficial if improved. Some possible examples could include:

[35] Farid 2021.

[36] https://gifct.org

[37] *The Tech Coalition Industry Classification System* 2022.

- Sharing of detection models
- Preventing the stripping of ML generation parameters from JPEG and PNG metadata, either preserving it during transcoding or recording it to assist with metrics and future investigations
- Signal sharing of hash data, model metadata and TTPs[38] of CG-CSAM generation networks
- Sharing of information regarding known bad actors

Sharing these signals between platforms and child safety research organizations via industry groups such as the Technology Coalition[39] would also be beneficial to keep participants apprised of the latest threats and techniques for detection and mitigation.

6 AI Ethics and safety by design

Safety by design encourages thoughtful development: rather than retrofitting safeguards after an issue has occurred, technology companies should be considering how to minimize threats and harms throughout the development process. For generative ML, this concept should be expanded to the entire lifecycle of ML/AI: develop, deploy and maintain. Each of these parts of the process include opportunities to prioritize child safety.

When considering development, some tactical steps that can be taken include:

- Remove harmful content from training data, e.g. hashing and matching the data against hash sets of known CSAM, or using classifiers and manual review
- Engage in red teaming sessions to pressure test particular themes and content, e.g. what prompts generate CG-CSAM
- Incorporate technical barriers to producing harmful content, e.g. biasing the model against outputting child nudity or sexual content involving children
- Be transparent with training sets (especially in the open source setting) such that collaborators can independently audit/assess the content for harmful content

When considering deployment:

- For cloud based, incorporate harmful content detection at the inputs and outputs of your system, e.g. detecting prompts intended to produce CSAM, and detecting CG-CSAM that may have been produced
- For open source, evaluate which platforms you allow to share your technology, e.g. determine if those platforms knowingly host models that generate harmful content
- For platforms that share models developed by other organizations and persons, evaluate which models you allow to be hosted on your platform,

[38] Johnson et al. 2016.
[39] https://www.technologycoalition.org

e.g. only host the models that have been developed with child safety in mind

- In all cases, pursue content provenance solutions that occur as part of development rather than as a optional post processing step, e.g. training a watermark into the decoder of the model itself, or releasing ML/AI solutions that can reliably predict the synthetic nature of the content

When considering maintenance:

- As newer models are developed and deployed with safety by design principles, remove access to historical models
- Proactively ensure synthetic content detection solutions are performant on the content generated by the newer models
- Actively collaborate with special interest groups to understand how your models are being misused
- For cloud based, include clear pathways to report violations to the proper governing authority
- Share back with the child safety ecosystem known hashes of CG-CSAM, and known inputs that produce harmful content discovered as a part of this process

While the merits of keeping DMs proprietary versus freely available are topics of debate, the open-source model of Stable Diffusion was unfortunately released without due care for the safety of the public. It was trained on adult content, shipped with an easily removable safety filter, and put in reach of anyone with a reasonably modern GPU. The effects of this will be with us for some time to come: all while a thriving community around Stable Diffusion 1.5 continues to advance its capabilities.

7 Planning for advances in generative ML

While we hope that the recommendations in this paper will help stem the proliferation of exploitative computer generated content, technology will advance and strategies to counter it will need to advance accordingly. Even with the best mitigations, there will arrive a point in the future where the cost of training an entire base diffusion model from scratch will come down to a range affordable by an individual. This will negate some of the recommended countermeasures, such as filtering out sexual and/or illicit content from training data or building watermarking into the model. To plan for this eventuality, much more will need to be done in the areas of detecting ML-generated content, technical collaboration and legal harmonization.

And while the world is already poorly prepared for the advent of realistic still imagery, realistic full-motion content is undoubtedly on the horizon. At this point, it is likely that attention will shift towards attempts to produce full-motion CG-CSAM. It is crucial that general purpose models to produce video content be trained, weighted and watermarked in such a way that it makes it as difficult as possible to use for this purpose. This can be a second chance to develop and deploy generative ML in as safe and ethical a manner as possible, learning from the mistakes of early generative models.

References

Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," arXiv: 1812.08008 [cs.CV].

Christensen, Larissa S., Dominique Moritz, and Ashley Pearson. 2021. "Psychological Perspectives of Virtual Child Sexual Abuse Material." *Sexuality & Culture* 25, no. 4 (August 1, 2021): 1353–65. https://doi.org/10.1007/s12119-021-09820-1.

Corvi, Riccardo, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2022. "On the detection of synthetic images generated by diffusion models," arXiv: 2211.00680 [cs.CV].

Dou, Liyun, Guorui Feng, and Zhenxing Qian. 2023. "Image Inpainting Anti-Forensics Network via Attention-Guided Hierarchical Reconstruction." *Symmetry* 15 (2). https://doi.org/10.3390/sym15020393.

U.S. v. Whorley, 386 F. Supp. 2d 693 (United States District Court, E.D. Virginia, Richmond Division 2005), https://casetext.com/case/us-v-whorley-5.

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

Federal Bureau of Investigation. 2023. "Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes," I-060523–PSA. https://www.ic3.gov/Media/Y2023/PSA230605.

Fernandez, Pierre, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. 2023. "The Stable Signature: Rooting Watermarks in Latent Diffusion Models," arXiv: 2303.15435 [cs.CV].

Gal, Rinon, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion," arXiv: 2208.01618 [cs.CV].

Guarnera, Luca, Oliver Giudice, and Sebastiano Battiato. 2023. "Level Up the Deepfake Detection: a Method to Effectively Discriminate Images Generated by GAN Architectures and Diffusion Models," arXiv: 2303.00608 [cs.CV].

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. "LoRA: Low-Rank Adaptation of Large Language Models," arXiv: 2106.09685 [cs.CL].

Johnson, Chris, Lee Badger, David Waltermire, Julie Snyder, and Clem Skorupka. 2016. *Guide to Cyber Threat Information Sharing.* Technical report NIST Special Publication (SP) 800-53, Rev. 4, Includes updates as of January 22, 2015. Gaithersburg, MD: National Institute of Standards and Technology, October. https://doi.org/10.6028/NIST.SP.800-150.

Ricker, Jonas, Simon Damm, Thorsten Holz, and Asja Fischer. 2023. “Towards the Detection of Diffusion Model Deepfakes,” https://openreview.net/forum?id=RZHdb7FnqlY.

Ruiz, Nataniel, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2022. “DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation.”

Stanley-Becker, Isaac, and Drew Harwell. 2023. “How a tiny company with few rules is making fake images go mainstream.” *Washington Post* (March 30, 2023). https://www.washingtonpost.com/technology/2023/03/30/midjourney-ai-image-generation-rules.

The Tech Coalition Industry Classification System. 2022. The Technology Coalition, July. https://www.technologycoalition.org/knowledge-hub/the-tech-coalition-industry-classification-system.

18 U.S. Code § 1466A, Stat. (Apr. 30, 2003), https://www.law.cornell.edu/uscode/text/18/1466A.

18 U.S. Code § 2252A, Stat. (Dec. 7, 2018), https://www.law.cornell.edu/uscode/text/18/2252A.

18 U.S. Code § 2256, Stat. (Dec. 7, 2018), https://www.law.cornell.edu/uscode/text/18/2256.

Vincent, James. 2022. “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.

Wang, Zhendong, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023. “DIRE for Diffusion-Generated Image Detection,” arXiv: 2303.09295 [cs.CV].

Wu, Haiwei, and Jiantao Zhou. 2022. “IID-Net: Image Inpainting Detection Network via Neural Architecture Search and Attention.” *IEEE Transactions on Circuits and Systems for Video Technology* 32 (3): 1172–85. https://doi.org/10.1109/TCSVT.2021.3075039.

Yang, Ling, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. “Diffusion Models: A Comprehensive Survey of Methods and Applications,” arXiv: 2209.00796 [cs.LG].

Zhang, Kevin Alex, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. “Robust Invisible Video Watermarking with Attention,” arXiv: 1909.01285 [cs.MM].

Zhang, Lvmin, and Maneesh Agrawala. 2023. “Adding Conditional Control to Text-to-Image Diffusion Models,” arXiv: 2302.05543 [cs.CV].

THORN

Stanford | Internet Observatory
Cyber Policy Center

***Thorn** is a nonprofit with a mission to build technology to defend children from sexual abuse. Founded in 2012, the organization creates products and programs to empower the platforms and people who have the ability to defend children.*

*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.*