(Mis)use of Nude Images in Machine Learning Research

Arshia Arya[3] **Princessa Cintaqia**[2] **Deepak Kumar**[3]
Allison McDonald[2] **Lucy Qin**[1] **Elissa M. Redmiles**[1]
[1]Georgetown University [2]Boston University [3]University of California, San Diego
{lucy.qin,elissa.redmiles}@georgetown.edu
{cintaqia,amcdon}@bu.edu {aarshia,kumarde}@ucsd.edu

1 Introduction

Nudity detection is a task that has been studied by researchers for decades [1]. For training, testing, and benchmarking nudity detection algorithms, researchers typically scrape images from the internet or use existing datasets of nude images. While this practice is common for assembling datasets for general image-recognition tasks, nude images are particularly sensitive. Images that were consensually shared on publicly-accessible forums (e.g., Reddit) or adult content platforms (e.g., PornHub, OnlyFans) were never intended to be used in research. Furthermore, publicly-accessible forums have been documented to host communities explicitly for the nonconsensual sharing of nude content [2]. Such sharing is a common form of image-based sexual abuse (IBSA) [3], which is a category of technology-facilitated sexual violence that includes the nonconsensual creation and distribution of intimate content. IBSA can lead to serious legal, emotional, employment, and relational consequences [4, 5], including clinical diagnoses of post-traumatic stress disorder, anxiety, and/or depression [5]. One of the most traumatizing aspects is that once an image has been distributed online, people lose control over how it is further spread and used. A victim-survivor shared with Bates et al. that, "I didn't have control over who they were distributed to [...] that they were used maliciously and without my consent, and in my name, that was the part that violated me the most" [5].

Our team is currently conducting research that investigates the use of nude datasets in machine learning and computer vision literature. By keyword searching for common ML tasks involving nudity (see Table A), we found a seed set of 2048 papers. While we are still processing the full results, we identify several ethical challenges based on our initial observations after reading dozens of these papers. In this provocation, we aim to raise questions for researchers considering work in this space to evaluate at the start of their projects and prior to dataset collection.

2 Prompts for Researchers

We acknowledge that ethical guidance for researchers in this space is lacking. We propose several concrete prompts for researchers considering working with nude image datasets and show how they align with existing ethical principles, using the 2024 NeurIPS Code of Ethics [6] as an example.

Do not include examples in the paper. Through our initial work, we have found an alarming number of published papers that include nude image examples. Some papers censor body parts while leaving the subject's face visible. In one example, researchers searched for, collected, and published "upskirt" images that the authors knew were created nonconsensually. The subjects did not consent to the creation of the images in the first place, much less having their images shared in a research publication. Particularly in cases where the images are identifying, we argue that publishing such images directly constitutes IBSA. Though researchers may feel that examples contextualize a paper's results, text descriptions or stylized depictions created by an artist would suffice. This aligns with the NeurIPS Ethics Guidelines, which state that "datasets should minimize the exposure of any personally identifiable information, unless informed consent from those individuals is provided to do so."

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Acknowledge the harms of collecting data nonconsensually. Scraping nude images from the internet, especially from social media platforms, will result in collecting some that were nonconsensually created or uploaded [7]. To avoid collecting nonconsensual content, researchers might instead scrape from adult content platforms (e.g., Pornhub, OnlyFans). This, however, is direct theft of labor and still constitutes IBSA. In fact, sex workers are often excluded from discussions around IBSA but may experience financial loss, risk being outed as a sex worker, and bear the mental health effects of experiencing IBSA [8, 9]. Neither approach aligns with the NeurIPS Code of Ethics *Consent* guideline, which states that "any paper that chooses to create a dataset with real data of real people should ask for the explicit consent of participants, or explain why they were unable to do so." For some tasks, like testing nudity in generated images, the only goal of collecting a nude dataset is to train a classifier. In these cases, using existing classifiers may suffice and will limit the number of images that are nonconsensually collected and used for research.

Do not redistribute the dataset. If a study produces a new dataset, redistributing the images without the consent of the subjects is a form of IBSA. Increasingly, in some jurisdictions such distribution would be illegal due to "revenge porn" laws [10]. For researchers who have already collected datasets of nude images, the dataset should be deleted after the study is complete.

Carefully consider researcher data handling practices. Researchers who store nude datasets must carefully establish security and access protocols for the data. In line with the NeurIPS Code of Ethics, researchers should "leverage privacy protocols, encryption and anonymization to reduce the risk of data leakage or theft," and take particular care with sensitive data. Senior researchers should also consider how they work with graduate students on such projects: some students may be uncomfortable working with potentially illegal data; others may require additional oversight to ensure they are responsibly handling the data.

Consider your use case. We encourage researchers interested in nudity detection to weigh the benefits of their research against the harms of nonconsensually collecting or using others' nude images. If researchers intend for their nudity detection tools to be used downstream by tech platforms, consider that larger platforms are already devoting internal resources to the task, may have their own data, and may have more nuanced definitions of nudity than those considered by researchers.

3 Conclusion: where do we go from here?

Organizational responsibility. When academic works containing nude images are published through large organizations like IEEE, ACM, or the Neural Information Processing Systems Foundation, these organizations gain revenue from nonconsensually distributed nude images. They therefore have a responsibility to check the content they host for nonconsensual nude images, even those in which faces or body parts are censored. Organizations could create voluntary commitments (akin to those made by tech platforms to combat IBSA [11]) to remove publications that currently contain nonconsensual nude images and create measures to prevent their publication in the future.

Generative AI. It may be tempting for researchers to suggest using generative AI to create nude images. However, generative AI models for nude images have themselves been created using nonconsensually collected images. Though unlikely, an image used in the training set may be reproduced as output. We also do not know how likely it is for an output to closely resemble the likeness of a living person regardless of whether their image was used in training. There is still much research needed to assess these and other risks. The datasets that researchers have nonconsensually collected and shared may themselves be used to train generative AI models for creating nude images. This is something that researchers should be mindful of when considering any requests for data sharing.

Community norms. Ultimately, we need increased awareness within the research community around the potential harms of collecting and distributing datasets containing nude imagery. Though researchers may have good intentions, their actions could have negative consequences on the individuals whose images they possess. Through our current work, we have not yet seen language that suggests that researchers are treating nude images with greater sensitivity than other types of data. We have not seen descriptions of how this data is stored, access policies, and security practices to ensure that it is not accidentally leaked. The broader research community must set new norms that regard nude images with greater sensitivity at every stage of the research pipeline, from authors to reviewers to the organizations that are currently complicit in publishing nonconsensual nude images.

References

[1] J. Cifuentes, A. L. Sandoval Orozco, and L. J. García Villalba, "A survey of artificial intelligence strategies for automatic detection of sexually explicit videos," *Multimedia Tools and Applications*, vol. 81, p. 3205–3222, Jan 2022.

[2] S. Hargreaves, *'I'm a Creep, I'm a Weirdo': Street Photography in the Service of the Male Gaze*. No. 3165345, Rochester, NY: Routledge: Routledge Studies in Surveillance book series, 2018.

[3] A. Powell, A. J. Scott, A. Flynn, and S. McCook, "Perpetration of image-based sexual abuse: Extent, nature and correlates in a multi-country sample," *Journal of Interpersonal Violence*, vol. 37, p. NP22864–NP22889, Dec 2022.

[4] C. McGlynn, K. Johnson, E. Rackley, N. Henry, N. Gavey, A. Flynn, and A. Powell, "'it's torture for the soul': The harms of image-based sexual abuse," *Social & Legal Studies*, vol. 30, p. 541–562, Aug 2021.

[5] S. Bates, "Revenge porn and mental health: A qualitative analysis of the mental health effects of revenge porn on female survivors," *Feminist Criminology*, vol. 12, p. 22–42, Jan 2017.

[6] NeurIPS, "Neurips code of ethics," 2024. `https://neurips.cc/public/EthicsGuidelines`.

[7] N. Henry and A. Flynn, "Image-based sexual abuse: Online distribution channels and illicit communities of support," *Violence Against Women*, vol. 25, p. 1932–1955, Dec 2019.

[8] Scarlett Redman and Camille Waring, "Visual Violence: Sex Worker Experiences of Image-Based Abuses," Feb. 2022. Publication Title: National Ugly Mugs.

[9] L. Qin, V. Hamilton, S. Wang, Y. Aydinalp, M. Scarlett, and E. M. Redmiles, ""Did They F***ing Consent to That?": Safer Digital Intimacy via Proactive Protection Against Image-Based Sexual Abuse," in *33rd USENIX Security Symposium (USENIX Security 24)*, (Philadelphia, PA), pp. 55–72, USENIX Association, Aug. 2024.

[10] C. C. R. Initiative, "Nonconsensual distribution of intimate images," 2024. `https://cybercivilrights.org/nonconsensual-distribution-of-intimate-images/`.

[11] The White House, "White House Announces New Private Sector Voluntary Commitments to Combat Image-Based Sexual Abuse," Sept. 2024.

A Appendix

keyword	# papers
"pornographic images" + "machine learning"	899
"nudity detection"	335
"NudeNet"	76
"safety filtering" + "machine learning"	135
"adult image detection"	296
"pornographic image detection"	307
total	2048

Table 1: A selection of our keywords and the number of hits on Google Scholar.