Stop the Nonconsensual Use of Nude Images in Research

Princessa Cintaqia
Boston University
cintaqia@bu.edu

Arshia Arya
UC San Diego
aarshia@ucsd.edu

Elissa M. Redmiles
Georgetown University
elissa.redmiles@georgetown.edu

Deepak Kumar
UC San Diego
kumarde@ucsd.edu

Allison McDonald*
Boston University
amcdon@bu.edu

Lucy Qin*
Georgetown University
lucy.qin@georgetown.edu

Abstract

In order to train, test, and evaluate nudity detection models, machine learning researchers typically rely on nude images scraped from the Internet. Our research finds that this content is collected and, in some cases, subsequently *distributed* by researchers without consent, leading to potential misuse and exacerbating harm against the subjects depicted. **This position paper argues that the distribution of nonconsensually collected nude images by researchers perpetuates image-based sexual abuse and that the machine learning community should stop the nonconsensual use of nude images in research.** To characterize the scope and nature of this problem, we conducted a systematic review of papers published in computing venues that collect and use nude images. Our results paint a grim reality: norms around the usage of nude images are sparse, leading to a litany of problematic practices like distributing and publishing nude images with uncensored faces, and intentionally collecting and sharing abusive content. We conclude with a call-to-action for publishing venues and a vision for research in nudity detection that balances user agency with concrete research objectives.

Content warning: this work discusses sexual violence, such as the harms of image-based sexual abuse (Section 1) and the inclusion of sexually violent imagery in datasets (Section 4).

1 Introduction

Nudity detection is a task that has been studied by researchers for decades [12]. 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, the scraping of nude images raises unique ethical concerns.

Nude images online take different forms. Some nude images may have been created and/or made public without the consent of the person depicted. Indeed, publicly-accessible forums have been documented to host communities explicitly for the nonconsensual sharing of nude content [28]. In many other cases, nude images were created *and* posted to a particular website by, or with the consent of, the image subject. However, even images that were consensually shared on publicly-accessible forums (e.g., Reddit) or adult content platforms (e.g., PornHub, OnlyFans) were intended for dissemination to a particular audience under a particular arrangement (e.g., for pay), not for use and/or redistribution in machine learning research.

*Both authors advised this study.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The nonconsensual creation or distribution of nude images is a form of sexual violence termed image-based sexual abuse (IBSA) [49], which encompasses the nonconsensual creation (e.g., "upskirting," "downblousing," "deepfakes") and/or nonconsensual distribution of intimate content, as well as threats to cause these harms [30]. IBSA can lead to serious downstream consequences similar to physical sexual abuse [41, 7]. According to some victim-survivors,[2] one of the most traumatizing aspects of IBSA is that once an image has been distributed online, the subject loses control over how it is further spread and used [7]. IBSA is also considered illegal in a growing number of countries (e.g., the United States [33], Canada [25], Mexico [2], South Korea [71], the United Kingdom [52]). Recent research has already documented the prevalence of known child sexual abuse material (CSAM) and nonconsensual intimate images of adults in the training datasets of major AI models as a result of unchecked scraping of images from the Internet [37, 62, 8, 9, 16].

Through a systematic analysis of 150 research papers on nudity detection, safety filtering, and related tasks, spanning 2002–2024, we document that these ethical challenges are almost entirely omitted from discussion in most research papers using nude images. We find that computer science research papers routinely engage in the nonconsensual collection and dissemination of nude images by scraping images extensively from the Internet, publishing those images as examples in research papers, and including them in publicly disseminated nudity detection models and datasets. In this paper, we identify and quantify common harmful data collection, distribution, and research practices.

Our goal is not to further stigmatize nudity and sexuality. We recognize that there are legitimate research needs for datasets of nude images. Rather, we argue that **research on nudity detection and related tasks must stop nonconsensually collecting and distributing nude images in order to avoid engaging in image-based sexual abuse.** We believe there are alternative paths forward that better embody our community's ethical standards and respect the dignity of all people.

2 Related Work

Our work focuses on the nonconsensual collection and use of nude images in research, but issues of consent with regards to data collection are not unique to nudity. Other scholars [19, 34, 8, 55, 20, 48] have critiqued the widespread practice of scraping data without consent to amass datasets for training and evaluating machine learning models. Previous work [55] found that among 125 machine learning datasets, only the authors of two datasets specifically asked and received consent from their subjects. Subjects are unlikely to be aware of what their data are being used for, and even if they are, it is common for their data to be used beyond what they initially consented to [3]. Furthermore, there are often no avenues for subjects to demand their data be removed from datasets, nor is this possible when datasets are shared and re-shared [55, 38]. Our work details how these same norms, when applied to nude images, perpetrate image-based sexual abuse. However, the harms that we elaborate on in this paper are symptom of "a culture where the appropriation of images of real people as raw material free for the taking has come be to perceived as the norm" [8].

We are not the first to address the harms of nonconsensually using nude images in machine learning research. Prior work has documented the existence of nonconsensually shared nude images in large image datasets (e.g., ImageNet [8], Moments in Time dataset [16], LAION-400M [9]) that are used for a wide variety of machine learning tasks. Most notably, researchers in 2016 conducted a review of 102 computer vision papers that focused on "pornography filtering." Their work surfaced the gender and sexual politics that are embedded in how computer vision researchers discuss and conceptualize their research (e.g., the common usage of female genitalia as a proxy for detecting "pornographic images" [22]). Our work, nearly a decade later, observes similar trends but focuses on the harms of research practices around the collection, distribution, and handling of datasets that are described to primarily contain nude imagery.

3 Methods

We systematically analyzed how researchers describe and use datasets containing nude images in computer science research. To do this, we collected computer science publications based on a set of search terms, manually annotated papers to extract relevant information for analysis, and then

[2]We use the term victim-survivor to capture the range of ways people who have experienced IBSA identify [59, 69, 30]. We recognize that some who have experienced harm may not identify with "victim" or "survivor".

conducted both quantitative and qualitative analyses. In total, we collected 150 papers after applying our inclusion criteria, which we manually annotated using a list of pre-defined questions. Additional details about our methodological process are captured in Appendix A.

Data Collection. We created an initial set of keywords based on exploratory review of highly-cited papers in nudity detection. We identified commonly used datasets and recurring task names associated with the use of datasets containing nude images (e.g., content filtering, adult image classification). Our final set of search terms can be found in Appendix A.1. We then used our search terms in Google Scholar, which searches the full text of manuscripts, and collected 1204 papers for further review. Our intention was to capture research whose technical tasks rely on nude image datasets. Our data collection therefore does not exhaustively cover all publications that use datasets containing nude images.

Inclusion Criteria. Our keyword search procedure returned 1204 papers. We focused our inquiry on (1) computer science publications that (2) use a dataset containing real (non-generated) nude images. We limited our analysis to computer science venues to allow for clearer comparisons in dataset framing, usage, and justification across papers that share technical assumptions, audiences, and publication norms. While the majority of these venues were related to AI/ML, 10 were related to security/privacy. However, almost all of the papers were centered on techniques from AI/ML. We filtered papers in our dataset to only publication venues indexed by DBLP [36], a database that catalogs computer science literature. The DBLP team provided us with a list of all of the venues they index. This filtering left us with a set of 379 papers. To address criteria (2), our team read through each paper and manually determined if the paper used any real nude images.

Dataset Descriptions. After applying our inclusion criteria, we arrived at a final set of 150 papers and found that there is urgency in addressing the issue of nonconsensual nude image collection and distribution in research. While our set of 150 included papers that were published between 2002-2024, the majority (76) were published in 2019 or later (see Figure 1). The use of datasets containing nude images is occurring across a multitude of venues. In total, the papers spanned 110 venues, including eight papers at six A* venues (based on CORE rankings [13]) such as CVPR and AAAI (see Appendix A.2 for full list). Furthermore, many papers were published through prominent publishing organizations such as IEEE (52 papers) and the ACM (10 papers). The papers in our inclusion set had collectively received 5,846 citations at the time of collection (Fall 2024), with the median number of citations being 10. This demonstrates that machine learning tasks involving the use of nude datasets are an active area of research.

We observed that nude images are used by researchers around the world. In total, the papers were authored by researchers at over 200 institutions across 42 countries, including the United States where 12 of the 13 institutions were R1 universities, such as Carnegie Mellon, Georgia Tech, Northeastern University, and UC Berkeley.

Annotation. We manually annotated all 150 papers that met our inclusion criteria using a list of pre-defined questions (see Appendix A.2.1). Each paper was annotated for (1) key details about the dataset (e.g., how many nude images were collected and used), (2) information about any published example images, and (3) research practices around data handling. We also defined a set of supplemental questions to investigate the framing and stated goals of research in our inclusion set (available in Appendix A.2.2). We captured this information using direct quotes from the papers. We randomly sampled and independently annotated papers from our inclusion set with these additional questions until we collectively reached thematic saturation [27]. This process included five papers published at A* venues. Since A* venues may be indicative of (and also set) norms for the rest of the research community, we identified and included 3 additional papers in our dataset that were published at A* venues. In total, **74 papers** were annotated with the supplemental framing questions.

Data Analysis. We analyzed the results from each annotation question using methods according to the data type. Most of our annotation questions required numerical or categorical responses (e.g., "How many nude images are in the dataset?") that we quantitatively analyzed to get descriptive statistics. To answer the annotation questions that required qualitative responses (e.g., "Does the paper discuss a social goal?"), we directly copied quotes from the papers. We then analyzed the quotes through an inductive coding process that is discussed in greater detail in Appendix A.3.

Ethical Considerations. Due to concerns about the inclusion of sample images in papers and some of the research practices we describe in our findings, we have taken measures to reduce the

discoverability of the papers within our corpus.[3] We intentionally do not share links to public datasets found during our analysis (though we make exceptions for well-known datasets, such as NudeNet). Instead of referencing specific papers, we have assigned each paper a random numerical ID. Papers published in A* venues are designated as PaperID^{A*}. We are in the process of responsibly disclosing the papers that contain uncensored nude images (with faces) to relevant publishers (e.g., IEEE, ACM). More details about this process and researcher safety are included in Appendix A.4.

4 Data Collection and Distribution Practices

Through our analysis, we observed a pattern of widespread nude image collection without the consent of the subjects depicted. Across the 150 papers in our annotated dataset, 126 unique datasets were used by researchers (including both existing and newly created datasets). In total, over 8 million images were collected and used. The largest dataset (Paper1052) contained 5,002,000 nude images and the smallest dataset consisted of 50 nude videos (Paper252). While 39 papers exclusively used previously collected datasets, almost 2 million images were newly collected by 98 papers that created their own datasets. Thirteen papers did not specify any details about their dataset.

4.1 Using nude images nonconsensually in research is harmful

Subjects' lack of consent is unacknowledged. Out of the 74 papers that we annotated using our full set of questions, we did not find any papers that discussed (the lack of) consent from image subjects. Furthermore, despite the real risk of emotional and physical harm that comes from redistributed nude content [29], none of the papers we analyzed explicitly considered the image subjects' safety in their decision to conduct the study—despite using this data to purportedly enhance the safety of the Internet (see Section 5.1). This lack of consideration for subjects is well captured in an essay by cultural worker Livia Foldes on the nude images published through the most popular pretrained nudity classifier, NudeNet, which noted that, "the agency of the people captured in [the dataset] has been so thoroughly denied that their consent, or lack thereof, was never mentioned by the researchers who used their intimate photographs to train an algorithm" [37].

Regardless of whether the content being scraped was originally created and/or distributed consensually, it is being collected nonconsensually by researchers for use cases that are unbeknownst to the data subjects. Studies on other social media platforms have found that users are largely unaware that their "public" posts could be used for research, and preferred to be asked for consent before their data was used [17]. Similarly, those who have voluntarily shared their nude images online likely did not anticipate the possibility of having them collected, stored, and redistributed by researchers. Even when a platform has formally authorized data access for research use (e.g., through an API or terms of service), researchers should not use a platform's terms of service as the arbiter of ethical data use and instead consider the safety and dignity of the data subjects across the benefits and context of the research.

Using commercial content is still harmful. To avoid collecting content that was nonconsensually created or shared, some researchers instead scraped from adult content platforms. In total, 18 papers stated that they collected data from adult content platforms while another 11 used a particular scraping tool that contained almost 4000 links to content on adult content sites. However, scraping data from adult content platforms is still a violation. Sex workers are often excluded from discussions around IBSA, yet they face the same mental health harms from experiencing IBSA, as well as the risk of being outed as a sex worker and the loss of livelihood from stolen content, among other threats [54, 50, 68]. Furthermore, while large adult content platforms have more robust content moderation practices, images shared for the purposes of perpetrating IBSA commonly end up on smaller adult content platforms, some of which have no mechanisms for victim-survivors to seek content removal [31].

Nude images are collected in excess. The lack of consideration of image subjects is further evident in the excess collection of new images. As mentioned, 98 of the papers in our annotated corpus created new datasets. This decision is rarely supported by additional reasoning. When justification was provided, it was short and vague. For example, Paper862 created a new dataset consisting of 18,000 images to have "diverse and challenging images." Paper789 chose to develop a new dataset of nearly

[3] We will make the dataset of papers available for researchers upon request on a case by case basis.

38,000 images, claiming that "there is no current dataset accessible for the classification of multiple body parts nudity." Meanwhile, Paper1100^{A*} collected "professional and amateur pornography" that was "intentionally recent" to avoid overlap with existing data.

Norms around collecting "public" data without consent are extractive, regardless of the type of data collected [43, 34]. However, the same practices applied to nude content can lead to even further harm and legal risk for researchers, necessitating higher standards for care and caution.

4.2 Researchers are amassing datasets of IBSA

Researchers are further amplifying harm by collecting and using abusive content. The vast majority (98) of papers we annotated created new datasets.

Nonconsensual intimate content is difficult to remove from the Internet. Scraping nude images from the Internet, especially social media platforms, will almost inevitably result in collecting some that were nonconsensually created or uploaded [29]. Perpetrators often nonconsensually distribute intimate content to social media platforms, online forums, and adult content sites. Once shared, the process of content removal is labor-intensive and can be re-traumatizing for victim-survivors. It may takes weeks for platforms to respond to their requests, if they respond at all [31]. In the meantime, victim-survivors repeatedly experience harm by spending hours submitting additional content removal requests and checking if the content is still public. Unmoderated adult content platforms may not offer any meaningful process for content removal. By collecting and using this data, researchers inadvertently increase its visibility and perpetrate harm by creating new avenues for dissemination.

Abuse material is incidentally collected. Of the 98 papers that created new datasets, the vast majority (60) did not mention a specific data source. Those that did reported collecting data from Reddit (Paper998, Paper802, Paper1100^{A*}), search engines including Google, Bing, and DuckDuckGo (Paper88, Paper777^{A*}, Paper815, Paper976), Tumblr (Paper795), "a large image search index (of a few billion images in size)" (Paper1038^{A*}), "public social networks or image sharing sites" (Paper 862), and 4chan (Paper773^{A*}). Such platforms are also commonly used for disseminating nonconsensual intimate content, sometimes in ways that victim-survivors are unaware of. Once discovered, victim-survivors often encounter difficulties with content removal [31]. Researchers might therefore be unintentionally collecting abusive content.

Abuse material is intentionally collected. In other instances, researchers intentionally included abusive content. Despite acknowledging that "it is illegal to possess and distribute [upskirt] images," Paper887 collected 1,637 upskirt images from "the Internet." Meanwhile Paper307 constructed a dataset that they described as being extracted from "hidden or self cameras." Paper6^{A*} uses a publicly-accessible dataset of URLs for 1,300,000 "NSFW images." Upon inspecting the URLs, we found that they pointed to images on specific subreddits. One of the folders intentionally included content depicting sexual violence, collected from two subreddits that have since been banned from the platform ("r/StruggleFucking" and "r/rape_roleplay").

Eighteen papers specifically mentioned collecting data from adult content platforms. In particular, Paper873 described collecting "unprofessional porn" that was "taken by unprofessional photographer including both nudity and sexual behaviour. The unprofessional porn images usually contain complex backgrounds and the image quality is poor." Based on the description alone, it is likely that the authors either collected stolen content from small-scale creators or collected content that was nonconsensually created (e.g., through a hidden camera). Similarly, Paper936 collected "amateur pornographic images." As previously mentioned, it is particularly challenging to remove nonconsensually distributed content from small, unmoderated adult sites since some do not provide any means of reporting content.

Our corpus also included papers (Paper1090, Paper806, Paper1083) that used child abuse material by partnering with law enforcement directly. These datasets were held by law enforcement, rather than the researchers, and had strict data handling protocols that prevented the researchers from ever viewing the content. This is a positive indication that this type of research is possible to do responsibly. However, not all papers did this. One paper (Paper815) described CSAM detection as a goal of their research but did not partner with law enforcement. Furthermore, they alluded to collecting and reusing nude images of children and minors with keywords such as "girls, boys, teenager..." in

conjunction with keywords related to sexual acts ("licking, lick, sucking, suck, blowjob, fellatio"). This likely constitutes CSAM and is illegal in most jurisdictions [18].

Data sources are largely undisclosed. We emphasize that the observations made above are based on a small number of data source disclosures (or through further investigation of datasets and third-party tools). Of the 98 papers that created new datasets, many did not state a data source (19) or included vague references to the "Internet" or "websites" with no specified source (60). Of the 38 papers that mentioned a specific online source, most descriptions were short and vague (e.g.,"samples from public social networks or image sharing sites" (Paper862)). By purposely obscuring the data sources or neglecting to include these details, researchers are evading accountability. Therefore, our observations are an under-reporting of the full extent of harm and the potential illegality of images that have already been collected within our corpus.

4.3 Images are distributed without consent via publication, annotation, and open science

We identify three primary pathways through which ML researchers further distribute nude images without the image subject's consent: publishing example images in research papers, sharing with annotators, and disseminating image datasets and models trained on those datasets. We emphasize that this further distribution of nude images *without consent* fundamentally constitutes IBSA.

Publishing example images. A vast majority (87) of papers in our corpus, including those published in A* venues (4), embedded example nude images in their papers. Nine papers did not censor the images at all and 28 papers only censored (e.g., applied a blur effect) the bodies of those depicted, while leaving faces visible. Over 816 images were included across 87 papers. Paper898 contained over 40 images (faces censored) while Paper815 contained over 10 images that were completely uncensored. Six papers, including Paper6^{A*}, included example images of explicit sexual acts. Paper887 published 21 upskirt images, despite acknowledging that such images are illegal to disseminate in some jurisdictions.

One of the primary challenges that victim-survivors of IBSA face is stopping further dissemination of nonconsensually created and/or shared content [67, 31]. The distribution of nude images as examples in academic papers contributes to this challenge. Example nude images are unnecessary for readers to understand nudity as a concept; text descriptions or artistic depictions would suffice.

Sharing with annotators. Many papers (44) described the need for manually labeling nude images, including annotating "centers of private parts" (Paper883, Paper1066, Paper1065) and specific sexual acts such as "Vaginal Penetration, Anal Penetration, BDSM, Bestiality" (Paper815). In total, researchers mentioned annotating over 800,000 images across 44 papers. Many of these papers were opaque about who performed annotations. For example, Paper789 had only two authors but "eighteen human annotators." Only four papers explicitly note that the researchers outsourced annotation work, such as Paper873, which "employ[ed] six person[s]" to annotate the dataset, and Paper978, which "invited 10 students in our laboratory to manually filter out the images," which led to a dataset of "nearly 150,000 pornographic images and about 500,000 normal images." Three of the four papers that outsourced annotation provide no details of annotation policies, tools involved in the annotation process, protections against dissemination of the nude images by annotators, or protections for annotators against viewing abusive content (given that some researchers collected abusive content).

Releasing datasets. Nude datasets serve as a concentrated repository of nonconsensually collected and distributed intimate content, making it more difficult for image subjects to seek removal of material depicting them from the Internet. Although there are norms within the ML community towards publicly sharing datasets, public accessibility is not appropriate in this context. We found three papers, including Paper777^{A*}, that made their datasets publicly available. Nude datasets from Paper777^{A*} and Paper862 are still publicly accessible at the time of this publication submission through HuggingFace and figshare. Furthermore, three papers made nude datasets available upon request. While three papers mentioned deciding not to distribute their dataset, decisions to limit or withhold access were not necessarily motivated by concerns about the privacy of those depicted. Paper998 expressed desire to make data public but was thwarted by existing laws: "Despite our every effort to make the dataset content and data acquisition process as transparent as possible, due to the nature of data and applicable laws the dataset cannot be shared publicly." Furthermore, Paper845 made the decision not to share the data because, "the images used in this paper... likely contain copyrighted content," rather than as an abuse or privacy consideration.

5 Research Practices

Established research ethics require that researchers decide what research practices are morally acceptable to achieve their aims. Prior work on ethics in computer security [35] highlights two core ethical frameworks that researchers can use to reason about their decision to engage in behavior that harms individuals. Here, we leverage both frameworks—consequentialist and deontological ethics—as lenses through which to consider the justifications for the research practices we observed. Further, we consider how existing mitigations against unethical research practice—institutional review boards and publishing-venue-mandated ethics statements—have failed to guard against harm.

5.1 Research objectives do not justify the means

Consequentialist ethics centers a tradeoff between the harms and the benefits of a decision. To justify such a tradeoff, the benefits of the decision must be clear *and* outweigh the harms. Of the 74 papers we annotated for framing, the vast majority (51) of those that specified any motivation (59) reported their aim as reducing the prevalence of pornography on the Internet. Most of these papers described the benefit of their work as combating the moral and societal ills of pornography without providing concrete examples or citations. For example, papers assert that "Internet pornography content affects many people's life especially adolescents and creates many social problems and moral issues" (Paper1079) and "The amount of digital pornographic content over the Internet grows daily and accessing such a content has become increasingly easier. Hence, there is a real need for mechanisms that can protect particularly-vulnerable audiences..." (Paper821). Yet, pornography is legal in a large number of jurisdictions, and scientific evidence about its harmfulness to adults is inconclusive at best, with recent meta-reviews suggesting that experiences of pornographic use as problematic are "actually more related to [individuals'] interpretations of that use rather than the use itself. Specifically, religious qualms...and moral disapproval of" their own use of pornography [26].

Further, not all nude images on the Internet are sexual or pornographic. People have a wide range of purposes for posting and consuming sexual and/or nude content, including to explore their sexuality, seek positive body validation from friends, or receive sexual education [47, 21, 50, 65, 39]. Indeed, condemning sexual content that is not abusive may in fact cause harm: prior work suggests that the censorship of sex and sexual speech on the Internet (i) increases stigma, therefore harming victim-survivors of image-based and other forms of sexual abuse, and (ii) decreases general access to sexual and reproductive health resources, which in turn plays a role in preventing in-person and online sexual abuse of both adults and children [66, 30, 70, 61]. Furthermore, there is hypocrisy in claiming the moral need to remove nude and sexual content from the Internet while creating repositories of such content and aiding their spread through the distribution mechanisms discussed in Section 4.3.

While more than half of the papers we annotated for framing purported to protect children from the harms of pornography, only six papers mentioned the issue of CSAM. Even fewer specifically worked on technical solutions to address this issue. This has echoes of the classic use of child protection to induce moral panic [40] and justify censorship, rather than an articulation of work on a substantive societal problem that could justify the consequences of harmful research practice.

5.2 Subjects are dehumanized

The other ethical framework highlighted by Kohno et al. [35], deontological ethics, at a high level suggests that "we have a duty to treat all other human beings... as 'ends and never purely as means.'" Under such a framework, there could be no justification that any societal benefit (an end) justifies using an individual depicted in an image as a "means" by using that person's nude images without consent (violating multiple of their rights: to privacy, autonomy, etc.). Given the language several papers use to refer to nude images, however, we hypothesize that the researchers may have so dehumanized the people depicted in the content that they have forgotten that they are people worthy of respect and protection. For example, papers described nude images and/or sexual content as "obscene" (Paper1079, Paper1095, Paper878, Paper39), "virus" (Paper1066), "offensive" (Paper897, Paper899), "harmful" (Paper1184^{A*}, Paper919) and "inappropriate" (Paper1187 and Paper6^{A*}).

5.3 Ethical considerations are overlooked

We specifically annotated papers in our corpus for discussions of ethical or privacy concerns and only found four papers (Paper6^{A*}, Paper773^{A*}, Paper853, Paper733) that made any mention of either. In these handful of instances, three papers focused on data access. Protective measures included limiting access to one individual (Paper853), not making data public (Paper6^{A*}), and not using crowdsourced/third-party annotators (Paper773^{A*}) (though the concern in this paper was to avoid exposing annotators to "disturbing and unsafe" images, rather than protecting the image subjects).

Out of the 150 papers we annotated, only eight discussed data security. For example, Paper733 and Paper806 mentioned using designated servers. Four papers (Paper733, Paper853, Paper724, Paper773^{A*}) limited dataset access to only authors of the papers to prevent "exposing harmful content to other people" (Paper724). Even then, none of the papers mentioned any deletion plans for the nude images they used. We urge researchers who have already collected datasets of nude images to remove them from any public repositories they are shared to and to securely delete images that are no longer being used.

Only two of the papers we annotated, Paper724 and Paper1108^{A*}, mentioned seeking (and receiving) IRB approval. Institutional review boards such as IRB or ERB offices must consider the potential harms and legality of collecting nude images from online spaces. Such boards should require researchers to articulate their data sources, construct detailed data security plans, and disclose any third-party data sharing (either direct sharing with other individuals and/or with third-party tools).

5.4 Classification boundaries lack thoughtful definition

Classification objectives are largely undefined. All but one paper we annotated for framing described their technical purpose as detecting nude, "adult," and/or "pornographic" images. Despite making claims such as, "exposure to the sea of pornography can lead to many social problems, including cyber-sex [addiction]. It is now an urgently necessary task to prevent people, especially children from accessing this type of harmful material" (Paper823), the papers we analyzed commonly did not articulate what they were trying to remove. More than half (39) of the fully annotated papers did not provide any specific criteria for classification. The majority of criteria (35) related to identifying body parts ranging from specific mentions of "anuses, female breasts, female genitals, male genitals" (Paper1108^{A*}) to vague references (e.g., "private body parts" (Paper1184^{A*}), "certain parts of the body" (Paper862)). The lack of specificity not only makes it difficult to understand the accuracy of model evaluation, but suggests they may capture entirely unrelated content. The lack of clear classification criteria establishes a norm that disregards censorship as an outcome.

Broad and underspecified classification criteria are also misaligned with the content moderation needs of online platforms, which must carefully navigate the balance between moderation and censorship [4, 24, 64, 46, 63]. Many papers that we fully annotated justified collecting and using nude datasets by suggesting that their work could be adopted by online platforms. Yet, what qualifies as nudity, in the context of content moderation, is nuanced, contested, and continually negotiated between users and large platforms. For example, Meta has revised their policy multiple times in response to user concerns about censorship and inequitable enforcement [42]. As a result, instances of nudity/sexual content related to health, education, and artistic expression are now excluded from moderation. In contrast, none of the fully annotated papers we reviewed accounted for context-specific exceptions, suggesting that researchers' criteria are disconnected from how content moderation occurs in practice.

Classification is fundamentally difficult. The delineation of what is nude/non-nude or sexual/not sexual is inherently subjective. There is a long history of debate in legal literature regarding these definitions [5], yet only one paper (Paper799) in our fully annotated corpus discussed defining the classification threshold as a limitation of their work. More than half (39) of the fully annotated papers lacked any specific criteria for classifying content as nude (or sexual) or not nude (or non-sexual), treating the threshold as well-understood and objective (e.g., "each image contains an obvious instance of pornography" (Paper1090)). In other cases, researchers made arbitrary classifications without justification or definitions, such as Paper799 which scored the "severity" of content (from less to more severe): "Female breast 1, Female buttock 2, male buttock, 2, Female genitalia posing 2, Female genitalia sexually active 3, Male genitalia 3, Sex toys 3, Coitus 4, Anal 4."

These classification and definitional decisions have large-scale societal consequences. Nudity classification algorithms disproportionately rate images of women as sexually suggestive [53]. For example, images of women athletes are more likely to be flagged as depicting nudity or being sexually suggestive than similar images of athletes who are men [23]. Indeed, a prior analysis of computer vision literature well identified the implicit assumption "that pornography is limited to images of naked women; that sexuality is largely comprised of men looking at naked women" [22]. While we cannot ascertain the gender of a subject based on an image, we observed that the example images in papers aligned with their analysis. We urge researchers to directly engage with the inherent subjectivity in classification and reflect on how their positionalities (e.g. personal values, politics, experiences, cultural values) affect the classification objectives they form [11, 56]. Furthermore, by creating a binary classifications of whether an image is "pornographic" or not, "voluntarily shared nudes, commercial porn, sex work, human trafficking, queer activism, NSFW fanart and paedophilia will converge, unhelpfully, into a muddy pit of obscenity" [63].

6 Discussion

Although datasets containing nude images have been used for well over two decades in computer science research, there has been no guidance on their usage aside from a relatively recent ban on the use of one specific image: a headshot collected nonconsensually from a Playboy centerfold in 1972 [44]. We observed virtually no difference in terms of how researchers handled nude images compared to other image and vision datasets. Despite dealing with highly sensitive data, papers largely make no effort to describe data storage, access policies, and security practices to ensure that the data is safe. More troubling, many papers themselves further disseminate nonconsensual images, and offer no—or very weak—justification for the harms of the research process. Here, we zoom out and discuss responses the computing community can take to avoid this harm in the future.

6.1 Publishing venues should step up to regulate and set norms

Frontier venues and major publishers such as NeurIPS, ACM, and IEEE set standards for the entire computing community—both in terms of scientific rigor as well as ethical norms. Over the last several years, most frontier venues have established research ethics guidelines that submitted papers are supposed to adhere to, and are increasingly defining the processes that reviewers should use to evaluate and potentially reject studies based on these guidelines. Existing principles, such as those around privacy [45, 32, 14, 1, 58], consent [45, 1, 14, 58], and minimizing harm [45, 32, 14, 1, 58], are sufficient to cover the practices we highlight here.

Of the eight papers published at $A*$ venues that we annotated, five were published in 2023 or later, by which time every venue had ethics guidelines already in place. This suggests that, while existing ethics guidelines cover principles that should regulate the use of nude images in research [6], reviewers are not raising concerns sufficiently to trigger a change in practice or additional discussion in the text of the papers. We urge reviewers to take additional care when reviewing papers that use nude images, especially related to data sources, dataset handling, appropriateness of the use cases, and the presence of example images.

Furthermore, while existing research ethics guidelines should be broad enough to cover harmful practices around nude image collection and use, publishing venues should take a stronger step to prevent harm by banning (and removing) the use of nude images in published manuscripts *where the subject is identifiable*. This, however, should not be taken as guidance for the wholesale banning of nude imagery in research publications, which may be relevant to the research in some cases.

6.2 Moving forward

As a research community, if we accept that there exist tasks for which highly sensitive and potentially harmful data are necessary, it is imperative that we also forge pathways to the ethical creation and handling of such data.

Research with a clear purpose. We call for researchers to carefully evaluate and clearly articulate the benefits of research that necessitates the collection and use of nude datasets. If the intention is for research to be adopted by law enforcement or industry, we urge researchers to partner directly with relevant stakeholders. This will ensure that design decisions are aligned with their real-world needs.

Building partnerships will also enable researchers to leverage existing data that stakeholders may already have access to (e.g., previously flagged content).

Opportunities and risks of AI-generated data. While this study examined papers using datasets of real people, our keyword search turned up many studies that were already using AI-generated nude images for various tasks. While generative AI might have the potential to be a less harmful strategy for creating a dataset of nude images, it is important to note that generative AI is not a harm-less approach. These models are most likely trained using nonconsensually collected and/or created images [62]. Researchers should take several things into consideration when deciding whether generated nude images are appropriate for their task. The risks and opportunities of this approach are still evolving. Indeed, members of our own research team hold differing stances on how to balance the possibility of harm reduction through AI-generated data and the possibility of furthering harm through the use of imperfect tools.

First, as discussed above, we encourage researchers to engage in their papers with how the goals of the research task outweigh the real costs of using generative models trained on nonconsensually collected and created content. Second, researchers should explicitly acknowledge in their manuscripts that current generative models are trained on the same data that researchers are trying to avoid using. Finally, researchers must think critically about the images they generate. For example, though unlikely, an image used in the training set may be re-produced as output. We also do not yet 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. More research is needed to assess these risks. We also note that generated images may not be suited for all research tasks. As such, generating images does not negate the need to develop a new, ethical model of data governance for real nude images.

The need for a new model of data governance. We call for a new model of data governance that centers ongoing, informed consent. Specifically, we propose a participatorily-governed data trust of nude and sexual images that are consensually collected for research purposes.

Prior work has shown that people are willing to donate highly sensitive data to research for altruistic reasons, and are more willing if they understand the purpose of the research [60, 51]. While further research is needed to understand the circumstances under which people would be willing to donate nude and sexual imagery to science, their willingness will also depend on their perceptions of the value of the research and the trustworthiness of the data trust. Beyond a mechanism to ensure that participants fully understand what it means for their data to be used for research, participants should be able to specify which types of tasks they are willing to participate in. For example, a participant may be willing to donate data for use in nudity detection, but wish to be excluded from any generative models.

While we believe such a data trust is feasible, it raises several challenges that will require future research. For example, a data trust will need careful measures to ensure that the donated data was created and donated consensually. One of the challenges we highlight above is that there is no way, from an image alone, to tell whether it was created consensually. Verifying consent will be critical for building a viable dataset. Furthermore, participatory governance by data subjects will require implementing complex access and use specifications and careful oversight of the ways that researchers use the data. A data trust would need careful data protection mechanisms, but its creation can include strong safety guarantees, for example by ensuring that a resulting generative model has not memorized them or that their likeness has not been generated.

Ultimately, while our research has uncovered many ways that the research community is in need of intervention on this problem, we believe there are pathways forward that better embody our ethical standards and protect the dignity of all people.

Acknowledgments and Disclosure of Funding

This work was partially funded by an unrestricted gift from Google. E. M. Redmiles was supported in part by NSF Awards #2344940 and #2513313. L. Qin was supported by the Fritz Fellowship through the Initiative for Technology and Society at Georgetown University.

References

[1] ACM. Acm code of ethics and professional conduct, 2018. https://www.acm.org/code-of-ethics.

[2] Anastasia Moloney. 'Revenge porn' victim fights back with Mexican law to stem digital violence. *Reuters*, November 2019.

[3] Adam J Andreotta, Nin Kirkham, and Marco Rizzi. Ai, big data, and the future of consent. *Ai & Society*, 37(4):1715–1728, 2022.

[4] Carolina Are. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 22(8):2002–2019, November 2022.

[5] Thomas C Arthur. The problems with pornography regulation: Lessons from history. *Emory LJ*, 68:867, 2018.

[6] Arshia Arya, Princessa Cintaqia, Deepak Kumar, McDonald Allison, Lucy Qin, and Elissa M Redmiles. (mis)use of nude images in machine learning research. In *Evaluating Evaluations: Examining Best Practices for Measuring Broader Impacts of Generative AI, a NeurIPS 2024 Workshop*, 2024.

[7] Samantha Bates. Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, 12(1):22–42, January 2017.

[8] Abeba Birhane and Vinay Uday Prabhu. Large Image Datasets: A Pyrrhic Win for Computer Vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1537–1547, January 2021.

[9] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[10] Steven Boada and contributors. scholarly: A library for retrieving author and publication information from google scholar. https://github.com/scholarly-python-package/scholarly, 2024. Accessed: 2024-05-12.

[11] Scott Allen Cambo and Darren Gergle. Model Positionality and Computational Reflexivity: Promoting Reflexivity in Data Science. In *CHI Conference on Human Factors in Computing Systems*, pages 1–19, New Orleans LA USA, April 2022. ACM.

[12] Jenny Cifuentes, Ana Lucila Sandoval Orozco, and Luis Javier García Villalba. A survey of artificial intelligence strategies for automatic detection of sexually explicit videos. *Multimedia Tools and Applications*, 81(3):3205–3222, Jan 2022.

[13] Computing Research and Education Association of Australasia (CORE). Core conference rankings portal, 2024.

[14] CVPR. Cvpr 2025 ethics guidelines for authors, 2025. https://cvpr.thecvf.com/Conferences/2025/EthicsGuidelines.

[15] Holly Else. How i scraped data from google scholar. *Nature*, 556(7700):273–274, 2018.

[16] Everest pipkin. On Lacework: watching an entire machine-learning dataset, July 2020.

[17] Casey Fiesler and Nicholas Proferes. "participant" perceptions of twitter research ethics. *Social Media + Society*, 4(1):2056305118763366, 2018.

[18] International Centre for Missing & Exploited Children. Child Pornography: Model Legislation & Global Review. https://www.icmec.org/wp-content/uploads/2016/02/Child-Pornography-Model-Law-8th-Ed-Final-linked.pdf, 2016. Accessed on 12/05/2025.

[19] Timnit Gebru and Remi Denton. Beyond Fairness in Computer Vision: A Holistic Approach to Mitigating Harms and Fostering Community-Rooted Computer Vision Research. *Foundations and Trends® in Computer Graphics and Vision*, 16(3):215–321, 2024.

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. Publisher: ACM New York, NY, USA.

[21] Christine Geeng, Jevan Huston, and Franziska Roesner. Usable Sexurity: Studying People's Concerns and Strategies When Sexting. In *The Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, 2020.

[22] Robert W. Gehl, Lucas Moyer-Horner, and Sara K. Yeo. Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science. *Television & New Media*, 18(6):529–547, September 2017.

[23] Gianluca Mauro and Hilke Schellmann. 'There is no standard': investigation finds AI algorithms objectify women's bodies. *The Guardian*, February 2023.

[24] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[25] Government of British Columbia. Intimate Images Protection Act, January 2025.

[26] Joshua B Grubbs and Shane W Kraus. Pornography use and psychological science: A call for consideration. *Current Directions in Psychological Science*, 30(1):68–75, 2021.

[27] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006.

[28] Stuart Hargreaves. 'i'm a creep, i'm a weirdo': Street photography in the service of the male gaze. In *Surveillance, Privacy and Public Space*, ssrn scholarly paper 10. Routledge: Routledge Studies in Surveillance book series, Rochester, NY, 2018.

[29] Nicola Henry and Asher Flynn. Image-Based Sexual Abuse: Online Distribution Channels and Illicit Communities of Support. *Violence Against Women*, 25(16):1932–1955, December 2019.

[30] Nicola Henry, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Adrian J. Scott. *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery*. Routledge, Abingdon, Oxon, 2022.

[31] Antoinette Raffaela Huber and Zara Ward. Non-consensual intimate image distribution: Nature, removal, and implications for the Online Safety Act. *European Journal of Criminology*, page 14773708241255821, July 2024.

[32] ICML. Icml pubilcation ethics, 2025. `https://icml.cc/Conferences/2025/PublicationEthics`.

[33] Cyber Civil Rights Initiative. Nonconsensual distribution of intimate images, 2024. `https://cybercivilrights.org/nonconsensual-distribution-of-intimate-images/`.

[34] Eun Seo Jo and Timnit Gebru. Lessons from archives: strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316, Barcelona Spain, January 2020. ACM.

[35] Tadayoshi Kohno, Yasemin Acar, and Wulf Loh. Ethical Frameworks and Computer Security Trolley Problems: Foundations for Conversations. In *USENIX Security*, 2023.

[36] Michael Ley. Dblp - computer science bibliography, 2024.

[37] Livia Foldes. NSFW Venus: From colonial archives to machine learning datasets, July 2024.

[38] Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. A framework for deprecating datasets: Standardizing documentation, identification, and communication. In *Proceedings of the 2022 acm conference on fairness, accountability, and transparency*, pages 199–212, 2022.

[39] Wendy G. Macdowall, David S. Reid, Ruth Lewis, Raquel Bosó Pérez, Kirstin R. Mitchell, Karen J. Maxwell, Clarissa Smith, Feona Attwood, Jo Gibbs, Bernie Hogan, Catherine H. Mercer, Pam Sonnenberg, Chris Bonell, and Natsal-4 Team. Sexting among British adults: A qualitative analysis of sexting as emotion work governed by 'feeling rules'. *Culture, Health & Sexuality*, pages 1–16, June 2022.

[40] Alice E Marwick. To catch a predator? the myspace moral panic. *First Monday*, 2008.

[41] Clare McGlynn, Kelly Johnson, Erika Rackley, Nicola Henry, Nicola Gavey, Asher Flynn, and Anastasia Powell. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, 30(4):541–562, 2021. _eprint: https://doi.org/10.1177/0964663920947791.

[42] Meta. Transparency Center, 2025.

[43] Surbhi Mittal, Kartik Thakral, Richa Singh, Mayank Vatsa, Tamar Glaser, Cristian Canton Ferrer, and Tal Hassner. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. *Nature Machine Intelligence*, 6(8):936–949, 2024.

[44] Dylan Mulvin. *Proxies: The cultural work of standing in*. MIT Press, 2021.

[45] NeurIPS. Neurips code of ethics, 2025. `https://neurips.cc/public/EthicsGuidelines`.

[46] Susanna Paasonen, Kylie Jarrett, and Ben Light. *NSFW: Sex, humor, and risk in social media*. Mit Press, 2019.

[47] Susanna Paasonen, Jenny Sundén, Katrin Tiidenberg, and Maria Vihlman. About Sex, Open-Mindedness, and Cinnamon Buns: Exploring Sexual Social Media. *Social Media + Society*, 9(1):20563051221147324, January 2023.

[48] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, November 2021.

[49] Anastasia Powell, Adrian J. Scott, Asher Flynn, and Sarah McCook. Perpetration of image-based sexual abuse: Extent, nature and correlates in a multi-country sample. *Journal of Interpersonal Violence*, 37(23–24):NP22864–NP22889, Dec 2022.

[50] Lucy Qin, Vaughn Hamilton, Sharon Wang, Yigit Aydinalp, Marin Scarlett, and Elissa 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)*, pages 55–72, Philadelphia, PA, August 2024. USENIX Association.

[51] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. Instagram data donation: A case study on collecting ecologically valid social media data for the purpose of adolescent online risk detection. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[52] Revenge Porn Helpline. Intimate image abuse laws in the UK.

[53] Piera Riccio, Thomas Hofmann, and Nuria Oliver. Exposed or erased: Algorithmic censorship of nudity in art. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

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

[55] Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Remi Denton, and Casey Fiesler. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–33, April 2023.

[56] Morgan Klaus Scheuerman, Allison Woodruff, and Jed R. Brubaker. How Data Workers Shape Datasets: The Role of Positionality in Data Collection and Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–42, October 2025.

[57] Seatgeek. Fuzzywuzzy: Fuzzy string matching in python, 2020.

[58] USENIX Security. Usenix security '25 ethics guidelines, 2025. `https://www.usenix.org/conference/usenixsecurity25/ethics-guidelines`.

[59] Sexual Assault Kit Initiative. Victim or Survivor: Terminology from Investigation Through Prosecution. Last accessed 2025-10-14.

[60] Anya Skatova and James Goulding. Psychology of personal data donation. *PLOS ONE*, 14(11):e0224240, Nov 2019.

[61] Zahra Stardust. *Indie porn: revolution, regulation, and resistance*. A camera obscura book. Duke University Press, Durham, 2024.

[62] David Thiel. Identifying and eliminating csam in generative ml training data and models. *Stanford Internet Observatory, Cyber Policy Center, December*, 23:3, 2023.

[63] Katrin Tiidenberg. Sex, power and platform governance. *Porn Studies*, 8(4):381–393, October 2021.

[64] Katrin Tiidenberg and Emily Van Der Nagel. *Sex and Social Media*. Emerald Publishing Limited, July 2020.

[65] Ari Ezra Waldman. Law, Privacy, and Online Dating: "Revenge Porn" in Gay Online Communities. *Law & Social Inquiry*, 44(04):987–1018, November 2019.

[66] Kate Walker and Emma Sleath. A systematic review of the current knowledge regarding revenge pornography and non-consensual sharing of sexually explicit media. *Aggression and violent behavior*, 36:9–24, 2017.

[67] Zara Ward. Revenge Porn Helpline Report 2022. Technical report, Revenge Porn Helpline, 2022.

[68] Josie Rachel West. Image-based sexual violence and imperfect victims: the case for platform cooperativism in the online sex industry. *Porn Studies*, pages 1–15, June 2024.

[69] Jessica Williamson and Kelly Serna. Reconsidering Forced Labels: Outcomes of Sexual Assault Survivors Versus Victims (and Those Who Choose Neither). *Violence Against Women*, 24(6):668–683, May 2018.

[70] UN Women, UNICEF, et al. *International technical guidance on sexuality education: an evidence-informed approach*. UNESCO Publishing, 2018.

[71] Sou Hee Yang. Unveiling technology-facilitated gender-based violence in south korea: Signs of gender-based violence, legal reforms, and the role of criminal law. In *(In) Visible Signs of Gender-Based Violence*, pages 353–382. Springer, 2025.

A Methodology (Additional Details)

A.1 Data Collection

We used the following search terms to collect our initial set of 1204 papers. As we identified common existing datasets of nude images, we also searched on those dataset names. We elide the dataset names here.

- "adult image detection"
- NudeNet
- "nudity detection"
- "pornographic image detection"
- "pornographic images" + "machine learning"
- "safety filtering" + "machine learning"

We searched for our selected list of keywords on Google Scholar, with each keyword or combination of keywords in quotes. We collected all pages of the search results using a Selenium-based web scraper. To address CAPTCHAs [15] and ensure representative data collection, we ran the data collection across multiple systems (and therefore IP addresses). We used `scholarly` [10] to collect structured metadata, such as venue, year, and author names. We manually annotated the affiliations of all the authors as this information was not present in metadata retrieved by scholarly. Lastly, we de-duplicated the data to remove overlapping papers that were collected across multiple search terms.

We matched the venues in our dataset to those indexed by DBLP by using strict matching and then used `fuzzywuzzy` [57] to match the remaining unmatched venues. Our goal in using fuzzy matching was to maximize recall in identifying papers published in computer science venues, even when venue names were formatted differently (e.g., "CVPR 2019" vs. "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition"). We selected token_sort_ratio as our matching algorithm because it handles word order differences effectively when comparing venue names. A match was considered valid when the similarity score was greater than or equal to 85. We then manually verified all fuzzy matches. This filtering resulted in a set of 379 papers that were published at DBLP-indexed venues.

A.2 Annotation

After inspecting each of the 379 papers and applying our inclusion criteria, we were left with 150 total included papers. Since the papers in our dataset include sensitive content that is potentially collected nonconsensually, such as identifiably nude example images, we will only make our dataset available upon request to minimize traffic to papers that distribute images and datasets.

We collectively reviewed an initial set of 20 papers to inform the development of our annotation framework. After discussing our initial observations, we drafted a preliminary set of annotation questions to guide systematic coding in the larger dataset. Our annotation questions are documented below and include the following topics: dataset details (e.g., data collection methodology, quantity of images in dataset), whether/how images from the dataset were published, the framing of the paper (e.g., its technical and social goals), and dataset handling (e.g., how researchers interacted with the dataset). Three of the authors used the questions to independently annotate the 20 papers. We then met as a team to discuss the process and made revisions to create a finalized set of annotation questions. We then manually annotated 150 papers that met the inclusion criteria. Among the 150 papers, 74 papers were annotated using supplemental questions concerning framing, social motivation, and terminology used. Papers were divided among three of the authors for independent deductive annotation until thematic saturation was reached, as previously noted in Section 3.

Eight papers were published at six A* venues: ACM International Conference on Multimedia (MM), Computer Vision and Pattern Recognition (CVPR), Annual ACM Conference on Computer and Communications Security (CCS), European Conference on Computer Vision (ECCV), USENIX Security, AAAI Conference on Artificial Intelligence (AAAI).

A.2.1 Annotation Questions

1. **Inclusion / exclusion check**
   *Answer Q1 only; complete the rest of the form **only** if the paper is included.*
   (a) Dataset contains real (non-generated) images that are only or primarily nude?
   ☐ Yes ☐ No
2. **Dataset details**
   (a) Does the paper create a new dataset or reuse a dataset from another paper?
   ☐ Create ☐ Reuse ☐ Both ☐ Other (_____)
   (b) If the dataset is reused, what dataset is it / what paper is it from?
   (c) If the dataset is new, where did they collect the images from? If they do not say, state "unspecified".
   (d) If the dataset is new, do they make it publicly available?
   ☐ Yes—freely online ☐ Yes—by request ☐ Not available ☐ Unspecified
   ☐ Other (_____)
   (e) What does the paper use the dataset to do? Be specific – e.g., train and test their classifier, fine-tune an existing model, etc.
   (f) How many nude images are in the dataset?
   (g) What do they describe about the characteristics of the dataset, if anything? (e.g., gender, age, ethnicity, image quality, lighting, etc)
   (h) What term does the paper use for describing the contents of the dataset? E.g.: nudes, explicit images, pornographic images
   (i) *Optional* Are there any other interesting datasets that they are using? E.g., datasets of bikini photos, the I2P text prompt dataset, a dataset of generated nudes, etc.
3. **Published images in the paper**
   (a) Does the paper include example images from the dataset? ☐ Yes ☐ No
   (b) If yes, how many images?
   (c) If yes, how are the images shown?
   ☐ Totally uncensored (body & face visible)
   ☐ Body parts censored, faces uncensored
   ☐ Fully censored (body parts & faces obscured/cropped)
   ☐ Other (_____)
   (d) *Optional* What other comments or observations do you have about the images visible in the paper? E.g. observed gender distribution, explicitness, etc.
4. **Dataset handling & researcher interaction**
   (a) Does the paper discuss data handling protocols for the data? E.g. storage procedures, deletion plans, etc. Use quotes if possible.
   (b) Does the paper describe any researcher interaction with the dataset? E.g. preprocessing to crop, manually annotating or inspecting annotations, etc. Use quotes if possible.
   (c) Do they name a funding source? Only specify if it's clear that the project was directly funded, not if they simply thank e.g. their university.

A.2.2 Additional Questions (paper goals and framing)

1. What is the technical application of the paper? E.g. nudity detection, internet filtering, etc.
2. Does the paper define nudity? State "no" or paste in a quote with the definition.
3. Who is the ultimate user of the tool, if stated? E.g. tech companies who want to detect nudity, ISPs looking to filter content, parents to install on their children's phones, etc. If not mentioned, state "unspecified". Use quotes if possible.
4. If mentioned, does the paper discuss a social goal? E.g., protecting children from nudity, purifying the internet, minimizing porn use at work, etc. Use quotes if possible.
5. Does the paper discuss any ethical considerations they made during the research? Use quotes if possible.

6. What terms does the paper use for describing sensitive aspects of the images? E.g.: genitalia, private parts, female breasts...

A.3 Analysis

First, the lead author read through annotations corresponding to questions that required qualitative analysis to develop initial codebooks per question. Afterwards, two of the authors independently coded the same set of 15 annotations (20%) for each of the seven questions using the initial codebooks. Disagreements were discussed, which led to further revision of the initial codebook. The rest of the annotations were then divided equally between the two researchers to be independently coded with the revised, final codebooks. The final codebooks are provided below in Appendix A.5.

A.4 Ethical Considerations

As we found that some papers distributed nude images by including identifiable examples (i.e., with faces visible) in the published manuscript, we decided to contact the relevant publishers. We see this as a form of responsible disclosure, and sought to determine whether it was possible to get some of the most egregious images removed from or censored in the manuscripts being hosted. When contacting publishers and their ethics committees, we emphasized that our concern is around the identifiability of image subjects and their lack of consent in having their images re-published in an archival database, rather than a complaint about the presence of nudity in research papers, which may be appropriate in some contexts. At the time of press, we have met with and/or reported images to 4 publishers. While internal review may have begun, no manuscripts have yet been modified.

At the start of the project, all researchers were aware that this project would entail reading and annotating papers that may contain nonconsensually shared nude images. We discussed the risks of engaging in this research. Annotators were reminded to take breaks and were provided access to a trained mental health clinician who they could contact at any time for support.

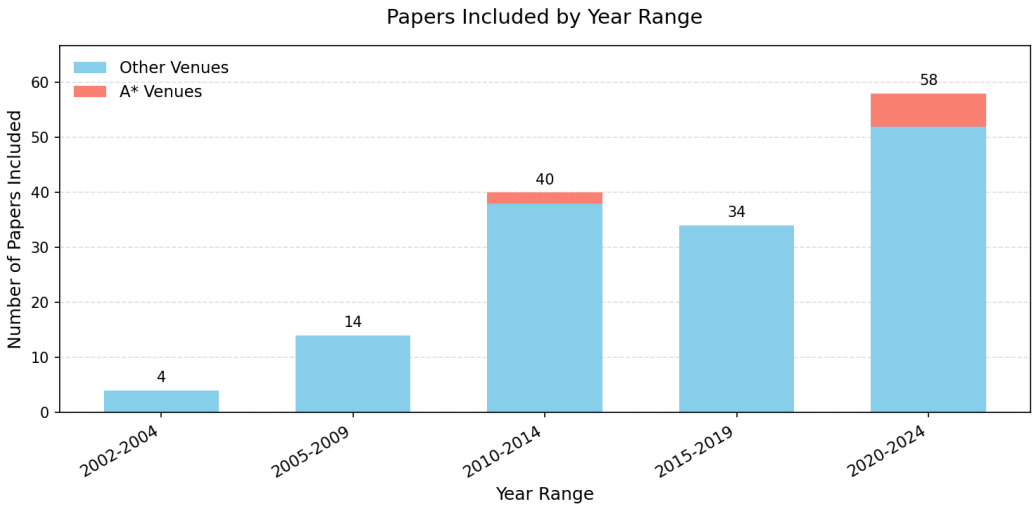

Figure 1: Number of papers included by 5-year range in the set of 150 included papers. The stacked bars differentiate between A* papers (pink) and other venues (blue). Note that there were no papers from the years 2000 or 2001 in our dataset.

A.5 Codebooks

Code	Definition
Does the paper define nudity? State "no" or paste in a quote with the definition.	
private_parts	Definition includes exposure of at least one of the "private parts of the body": "female breasts", "vagina", "penis", and "buttocks"

Code	Definition
naked	Definition includes naked bodies, without any clothes being present
skin	Definition describes large percentages of skin pixel or skin region
sex	Definition includes presence of sexual activities/scenes
female	Definition specifically mentions "female" body parts
male	Definition specifically mentions "male" body parts
unspecified	No definition of stated
Who is the ultimate user of the tool, if stated?	
companies	The tool is purportedly for companies (e.g., social media, video sharing, gaming)
parents	The tool is built for parents, mainly to protect their children from nudity online
children	The tool is built for children to protect themselves from nudity online (nudity ads, etc)
lea	The tool is built for Law Enforcement Agencies (LEA) to automatically detect CSAM and prosecute perpetrators who share CSAM
researchers	The tool is built to help researchers/AI practitioners to help with research or future work regarding nudity detection.
government	The tool purports to help the government identify content.
content_mod	The tool is purports to help human content moderators
unspecified	No end user specified
If mentioned, does the paper discuss a social goal of the paper? E.g., protecting children from nudity, purifying the internet, minimizing porn use at work, etc. Use quotes if possible.	
protect_minors	The social goal of the paper is to protect minors (children, teenagers, juveniles)
protect_society	The social goal of the paper is to protect society
protect_annotators	The social goal of the paper is to protect human content moderators or law enforcement agency staff from having to manually annotate content
purify_internet	The social goal of the paper is to "purify the internet"
address_csam	The social goal of the paper is to address CSAM
unspecified	The social goal of the paper is unspecified
Does the paper discuss any ethical considerations they made during the research? Use quotes if possible.	
not_public	dataset cannot be made accessible
limited_access	access limitations were made
beneficial	benefit surpasses harm
unspecified	no ethical considerations mentioned
other	
Does the paper describe any researcher interaction with the dataset? E.g. preprocessing to crop, manually annotating or inspecting annotations, etc. Use quotes if possible.	
annotate_nudes	manual annotation/labeling to determine nudes, sexual, and non-nude images and divide them into separate classes
annotate_body	manual annotation/labeling to select specific body parts
annotate_sex	manual annotation/labeling to select specific "sexual acts"/"positions"/"postures"
for_children	manually removing images inappropriate for children
choose_images	manually choosing nude images for created dataset/training set
crop_useless	manually cropping images to remove unnecessary parts/feature (e.g. background, logo. etc)
inspect	inspecting annotation, dataset flaws/mistakes, results/mistakes
none	no researcher interaction
unspecified	unspecified