
# Towards Targeted Obfuscation of Adversarial Unsafe Images using Reconstruction and Counterfactual Super Region Attribution Explainability

Mazal Bethany, Andrew Seong, Samuel Henrique Silva,
Nicole Beebe, Nishant Vishwamitra, and Peyman Najafirad,
*The University of Texas at San Antonio*

https://www.usenix.org/conference/usenixsecurity23/presentation/bethany

This paper is included in the Proceedings of the
32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the
32nd USENIX Security Symposium
is sponsored by USENIX.

# Towards Targeted Obfuscation of Adversarial Unsafe Images using Reconstruction and Counterfactual Super Region Attribution Explainability

Mazal Bethany, Andrew Seong, Samuel Henrique Silva,
Nicole Beebe, Nishant Vishwamitra, Peyman Najafirad *
*The University of Texas at San Antonio*

## Abstract

Online Social Networks (OSNs) are increasingly used by perpetrators to harass their targets via the exchange of unsafe images. Furthermore, perpetrators have resorted to using advanced techniques like adversarial attacks to evade the detection of such images. To defend against this threat, OSNs use AI/ML-based detectors to flag unsafe images. However, these detectors cannot explain the regions of unsafe content for the obfuscation and inspection of such regions, and are also critically vulnerable to adversarial attacks that fool their detection. In this work, we first conduct an in-depth investigation into state-of-the-art explanation techniques and commercially-available unsafe image detectors and find that they are severely deficient against adversarial unsafe images. To address these deficiencies we design a new system that performs targeted obfuscation of unsafe adversarial images on social media using reconstruction to remove adversarial perturbations and counterfactual super region attribution explainability to explain unsafe image segments, and created a prototype called uGuard. We demonstrate the effectiveness of our system with a large-scale evaluation on three common unsafe images: Sexually Explicit, Cyberbullying, and Self-Harm. Our evaluations of uGuard on more than 64,000 real-world unsafe OSN images, and unsafe images found in the wild such as sexually explicit celebrity deepfakes and self-harm images show that it significantly neutralizes the threat of adversarial unsafe images, by safely obfuscating 91.47% of such images.

**Disclaimer**. This manuscript contains harmful image content, such as sexually explicit, cyberbullying, and self-harm images that are highly offensive and might disturb the readers.

## 1 Introduction

OSNs have become an integral part of communication for many Internet users [3, 16]. But the ability to share content, especially images on these platforms leaves users vulnerable to unsafe content uploaded by bad actors. For example, recent studies suggest that women are increasingly experiencing image-based sexual abuse online [13]. Alarming incidents of cyberbullying involving images among teenagers have only increased [25, 39]. Furthermore, images depicting self-harm have recently soared on OSNs such as Instagram, and in 2018 alone, teens reportedly posted between 58,000 and 68,000 images with self-harm related hashtags on Instagram [55]. The sharing of such images that depict unsafe content such as sexually explicit content, images of cyberbullying and harassment, and images glorifying self-harm has emerged as a critical problem plaguing OSNs.

Faced with the serious threat posed by unsafe images, OSN platforms such as Facebook [20] and Snapchat [74] have deployed Artificial Intelligence/Machine Learning (AI/ML)-based detectors that can flag down such images. While these detectors are reportedly effective [53], they currently suffer from two key issues. *First*, these detectors lack the ability to explain the regions that are harmful in unsafe images. Explaining these regions is critically important to automatically obfuscate these regions so that OSN users are not exposed to such content [49, 50]. Moreover, explanation of such regions is also crucial for the inspection and analysis of these images by human moderators [20, 72] and law enforcement personnel [47]. However, explaining such regions is a major challenge since most explanation techniques [63, 85] are geared towards pinpointing only some important pixels in an unrelated manner near objects in images. But for explanation of unsafe regions for their obfuscation and inspection, pinpointing of meaningful segments (*e.g.*, private body parts in sexually explicit images) is needed. Thus, new explanation methods for explaining unsafe images based on meaningful image segments need to be developed. *Second*, although defenses against adversarial attacks against general AI/ML models have been proposed [31, 52], the threat posed by adversarial unsafe images, i.e., unsafe images that have been adversarially perturbed, has not been mitigated. As pointed out in recent studies [18], studying the adversarial influence of such detectors that operate in a hostile environment is of key importance

*Corresponding Author

for their real-world use [19, 61]. Adversarial unsafe images pose an important challenge to existing defense techniques since adversarial perturbations need to be removed from such images for further analysis of unsafe content in them. However, the existing defenses can only detect if images have been perturbed, but cannot remove such perturbations from images. New defenses that can remove these perturbations for further analysis of such images need to be formulated.

In this work, we report the first large-scale study on the critical real-world threat posed by adversarial unsafe images. We broadly address the following three research questions in our work: (1) Can we make an unsafe image a safer image? (2) Can we identify a specific area in the image that makes it unsafe? (3) Can a defense mechanism be established to stop an adversarial attack from misidentifying the area of an image that makes it unsafe? We first carry out an investigation into the capability of 3 state-of-the-art explanation techniques [63, 68, 85] to explain harmful content in unsafe images, and find that these techniques are extremely unsuitable in explaining unsafe images since they imprecisely provide scattered or sparse pixels as explanations of such images. We then carry out a large-scale experiment to study and measure the capability of 5 existing commercially available offensive image detectors [4, 6–8, 10] against adversarial unsafe images and find that all these detectors are vulnerable to adversarial unsafe images (*e.g.*, over 60% of adversarial unsafe images were able to fully evade detection).We then propose a novel system called uGuard to defend against the threat of adversarial unsafe images. uGuard uses a novel algorithm called Counterfactual Super Region Attribution (CSRA) that explains harmful segments in unsafe images, by optimizing the attribution of image regions pointed out by gradient-based methods against image segments in the unsafe image, while minimizing the area of the explanation region, and a novel image reconstruction approach called Adaptive Clustering for robust Semantic Representation (ACSR) that learns the distributions of both high and low-frequency signals in an image, and then removes adversarial perturbations from unsafe images by reconstructing the high-frequency signals in the image from a learned distribution of high-frequency signals of unperturbed images. Our evaluation of uGuard on 3 categories of unsafe images show that it is able to successfully reconstruct 90.94% of unsafe images, and reduce the risk of exposure of 96.94% of unsafe images. We run uGuard on 2 categories of unsafe images that we found in the wild (*i.e.*, sexually-explicit celebrity deepfake images on 4chan [11] and self-harm images from archive of an extremist OSN BestGore [5]) and found that our system neutralized the threat of 91.47% of these adversarial unsafe images found in the wild.

Our work makes the following contributions:

- We investigate the capability of state-of-the-art explanation techniques and commercially-available unsafe image detectors against adversarial unsafe images to provide an in-depth understanding of their vulnerabilities against this threat. Our analysis of 3 explanation techniques and 5 commercial detectors reveals alarming gaps in these technologies against adversarial unsafe images.
- To defend against the threat of adversarial unsafe images, we design a new system uGuard, which uses novel explanation and image reconstruction algorithms to make adversarial unsafe images safer. Our system has been publicly released to promote further research into adversarial unsafe images. [1]
- We evaluate uGuard on 3 categories of unsafe images [43, 76], including a novel dataset of self-harm images, and also run our system on 2 categories of unsafe images found in the wild. Our evaluation shows that uGuard is highly effective in obfuscating the harmful regions in unsafe images, reconstructing adversarial unsafe images, and neutralizing the threat of adversarial unsafe images found in the wild.

## 2 Background

Content obfuscation has been widely studied to enhance image privacy in various contexts [78]. Li et al. [50] show that image obfuscation can be an effective technique to hide the identities of individuals in a social media setting. One of the most popular approaches to content obfuscation are blurring methods such as Gaussian blur [40]. Another common content obfuscation approach is pixelization, where the original pixels are replaced by a smaller number of larger pixels [29]. A more intrusive obfuscation technique is masking. Masking usually involves replacing the image content with a uniformly colored rectangle [46]. Korshunov et al. [46] show that the masking technique is the most effective for hiding content, but the experiments of Li et al. [50] show that masking provides the least satisfaction out of the eight obfuscation techniques that they studied. Obfuscation is an established method of content control in social media settings, which could make it an important component in a system to protect users from unsafe content.

The current literature contains many works that detect unsafe image content, however, most of these works only focus on one category of unsafe content. The task of detecting sexually explicit images was achieved with high accuracy by Jin et al. [42] by modeling the problem as a multiple instance learning problem. The approach by Negri et al. [58] combines crowdsourced information alongside deep learning models to detect sexually explicit image content. On the other hand, the work of Nguyen et al. [59] use masking and CNNs to detect sexually explicit images. Vishwamitra et al. [76] approached the task of identifying cyberbullying image content by taking a multimodal approach that considers the input image as well as visual factors such as body-pose, facial emotion, object, gesture, and social factors. The study by Housseinmardi et

[1] https://github.com/SecureAIAutonomyLab/uGuard

al. [41] focused on cyberbullying images on Instagram and their correlation with other social media metadata. The work of Wang et al. [79] approached the detection of self-harm content on social media by analyzing image content in conjunction with the associated text captions and comments. On the other hand, works like Scherr et al. [66] only use image content to detect self-harm on social media using an AlexNet architecture. The approach by Xian et al. [80] aids their detection of self-harm images with the use of weakly supervised object detection techniques.

Adversarial attacks [23, 31] have been known to compromise AI/ML models, specifically vision-based models. While some works explore the vulnerability of existing AI/ML systems [33, 60], these works focus on general domains such as object detection. However, the real-world implications of adversarial attacks on existing safety and security critical systems is an area that needs more attention. For example, emerging studies have shown how adversarial attacks [22, 45] crafted to attack vision-based models in autonomous vehicles are capable of compromising them. The vulnerability of unsafe image detectors in adversarial settings however, is a safety-critical topic that has received significantly limited attention.

There are many techniques that have been developed to defend against adversarial attacks. According to Silva et al. [70], the four most common approaches towards the goal of defense involve: (1) modifying an AI/ML model, (2) training the model against adversarial examples, (3) transforming the input to reduce the impact of the adversarial perturbations, or (4) adversarial example detection. The first two methods require those who build the model to be aware of the threats of adversarial examples during their training or construction of their model. Examples of modifying the network include approaches such as Feature Squeezing [81], where the search space that is available to an adversary is reduced. Adversarial training was a concept introduced by Madry et al. [52] that has a model trained on adversarial examples to learn the features of specific attacks. On the other hand, the last two methods for achieving adversarial robustness use models that are separate from the classifier to manage adversarial examples. Input transformation defenses can range from compression techniques such as JPEG compression [28], to image reconstruction techniques such as Neural Representation Purifier (NRP) [56] and Adaptive Clustering of Robust Semantic Representations (ACSR) [69] by which an image is to be reconstructed without adversarial perturbation, though such methods may be dataset specific. Adversarial example detection methods [83] detect inputs that have been adversarially perturbed by finding outliers or by using neural networks that can distinguish between attacked and unattacked (clean) inputs. Although there has been much progress on fundamental works in adversarial robustness, there has been very little work on the application of this progress to unsafe social media content. Furthermore, the majority of work on adversarial attacks and defenses are evaluated on CIFAR-10, CIFAR-100, ImageNet, or MNIST datasets, leaving a gap in the current literature on adversarial robustness on unsafe social media content.

## 3 Threat Model and Problem Scope

**Threat Model**: We consider three types of social media users: (1) Perpetrators who create, store, or share adversarially unsafe images, (2) Target users who unwillingly receive or are depicted in unsafe images (3) and personnel who review unsafe images (such as OSN content moderators and law enforcement agents). The adversaries take advantage of open source methods of adversarially perturbing unsafe images to evade the automated unsafe content detectors. Our paper considers three types of unsafe content: sexually explicit, cyberbullying, and self-harm. We focus on these three image categories in this work based on the following reasons: (1) These categories of images are prohibited by popular social media platforms [75]. Social media platform guidelines are designed to maintain a safe and positive user experience, prevent harm, and comply with legal requirements [34]. (2) These types of images are of critical concerns to social media users pertaining to their safety and the large amount of traffic in these categories, informed to us by representatives from federal agencies we are working with. These content categories have been identified by federal agencies through various external reports such as [57, 67] and through correspondences with a member of a federal agency (who is also a collaborator in this work). (3) The availability of such datasets in existing literature. These image categories can be unsafe, yet are accessible for study. On the other hand, CSAM data is not made available for many ethical and legal reasons [65]. We only consider images, and not any text, user information or metadata.

In our system, we make the following assumptions: (1) the types of adversarial attacks are known to our system, and (2) the categories of unsafe images are known to our system. uGuard is applicable to unsafe images where specific regions are the causes of the image to be unsafe, e.g., the genitalia regions in sexually explicit images. uGuard may not be applicable to unsafe images that are not region-based, such as hateful memes, where the reason for the image to be unsafe is a combination of the image content and text overlaid on the image, or screenshots of hateful text.

## 4 Investigating the Threat of Unsafe Images

### 4.1 Investigating Explanation Techniques for Obfuscating Unsafe Images

In social media content moderation, human moderators are a core component of the image review process and the repeated exposure of harmful content to moderators has been acknowledged by courts to have caused harm [2]. Further-

more, in our discussions with law enforcement personnel, we learned that investigators who review images are often exposed to extremely unsafe image content during investigations. In both of these cases, there is a need to obfuscate the harmful part of the image to protect the image reviewer, while also maximizing the safe parts of the image that could contain vital information that is crucial evidence in the investigation process (e.g. identifiers of the people in the image, people's location, and age estimation in child abuse imagery). For example, in CSAM images, age estimation does not necessarily require the visibility of the unsafe image portions since age estimation techniques can use facial features, body proportions, alongside other traits to estimate age [65]. For these tasks, the targeted obfuscation of the unsafe regions of an image limit the harm faced by these personnel, while preserving the important contextual information in the safe parts to allow these personnel to get the information that they require from the image is needed. AI explainability has been previously used to improve the effectiveness of real-world safety-critical applications [14, 15] in multiple innovative ways. In the area of security of image sharing in OSNs, existing works [49, 50] have shown how region-based obfuscation of sensitive images can significantly mitigate the harmful effects caused by such content. Since OSNs use AI models [20, 74] to moderate unsafe images, we wanted to explore how explanations from these models can be used to obfuscate unsafe image content, since these models make predictions based on those regions.

Our goal was to investigate whether existing AI explanation approaches can be used to obfuscate only the unsafe regions in images, while preserving the rest of the information in those images. We conducted a preliminary experiment, by considering state-of-the-art image explainability methods and unsafe images consisting of sexually explicit images [43], cyberbullying images [76], and self-harm images which we collected (Further details about datasets can be found in Section 5). We used three explainability methods in our experiment: Grad-CAM [68], since it is representative of the explainability methods that rely on class activation maps to generate explanations, Integrated Gradients [73] since it is representative of explanation methods that output sparse pixels as explanations, and LIME [63] since it is representative of perturbation-based image explainability methods. We trained three binary ResNet-50 [38] classifiers to distinguish between safe and unsafe images from these three unsafe image classes. Then, we used the three explanation techniques to automatically obfuscate the unsafe regions in the cyberbullying images, pointed out by the generated explanations. For Grad-CAM and Integrated Gradients, we considered the top 20% of pixels identified for contributing to the model's decision. For LIME, we considered the regions identified as contributing to the model's decision. To visualize the ability of Grad-CAM, Integrated Gradients, and LIME to localize unsafe content in the image, we white-out the pixels or regions identified by these explainability methods. The results of obfuscating unsafe regions based on the three explanation techniques on a randomly selected cyberbullying image is depicted in Figure 1.

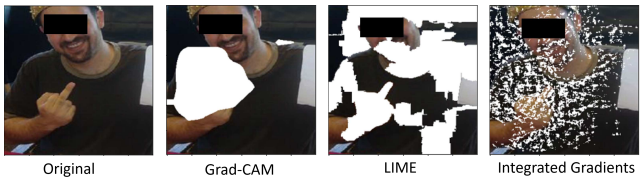

Figure 1: Samples of an unsafe image obfuscated according to the regions pointed-out by three explainability methods.

We found that none of the existing explanation techniques were suitable for automatically obfuscating the harmful regions in unsafe images. Grad-CAM imprecisely obfuscated an excessive region of the image besides the unsafe regions, resulting in lost information that is originally safe to view. LIME generated an explanation that resulted in excessive obfuscation consisting of scattered, unrelated pixels. Integrated Gradients on the other hand produced sparse pixel level explanations, which were unsuitable for targeted obfuscation. Furthermore, we conducted a quantitative analysis of these methods, presented in Table 1 to study the percentage of predictions changed and the percentage of the image that was obfuscated. From Table 1, Integrated Gradients was found to be significantly limited in masking the unsafe content (*i.e.*, low % of pred. changed), LIME obfuscated all of the unsafe content, but also obfuscated the safe parts of the image (*i.e.*, high % of image obfuscated), and Grad-CAM imprecisely obfuscated the harmful parts of the image (*i.e.*, missing multiple unsafe regions in sexually explicit images).

Our preliminary experiment indicates that state-of-the-art explanation methods are not suitable for targeted obfuscation of unsafe images, and that new explanation methods that are specific for explaining unsafe images that can find optimal unsafe regions to obfuscate while preserving as much safe information as possible are needed.

## 4.2 Evading State-of-the-Art Unsafe Image Detection

Recently, OSNs and other online platforms have increasingly deployed AI/ML-based automatic detectors to flag unsafe content. These detectors are comprised of advanced deep-learning based models that have demonstrated effectiveness in detecting such content. But at the same time, these models can also be compromised using adversarial examples [31, 52].

| | Grad-CAM | | Integrated Gradients | | LIME | |
|---|---|---|---|---|---|---|
| | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. |
| Sexually Explicit | 43 | 20 | 32 | 20 | 100 | 65.21 |
| Cyberbullying | 79 | 20 | 29 | 20 | 100 | 63.84 |
| Self-Harm | 65 | 20 | 41 | 20 | 100 | 71.92 |

Table 1: Experiment showing the unsuitability of different types of explanation methods for content obfuscation.

| Attack | State-of-the-Art Unsafe Image Detectors | | | | |
|---|---|---|---|---|---|
| | Clarifai (%) | Yahoo Open NSFW (%) | Amazon Rekog-nition (%) | MS Azure (%) | Google Safe-Search (%) |
| No Attack | 80 | 84 | 90 | 96 | 90 |
| Square | 22 | 6 | 50 | 68 | 76 |
| Square+GB | 4 | 4 | 74 | 94 | 76 |
| AutoAttack | 22 | 56 | 84 | 90 | 88 |

Table 2: Measuring the capabilities of state-of-the-art unsafe image detectors against adversarial unsafe images.

For example, recent works [22, 45] have demonstrated how vision-based models installed on self-driving cars can be compromised with adversarial attacks, rendering them unsafe. However, the real-world safety and robustness of unsafe content detectors against such adversarial examples is not a well researched area. Other works [18, 19], have pointed out how AI-based systems in security-critical applications must defend against adversaries that specifically target the system and will search for and exploit weaknesses for evasion or manipulation. Moreover, removal of such adversarial perturbations from unsafe images is even more critical, since the presence of such adversarial perturbations would render any further analysis (e.g., obfuscation of harmful regions) useless. To understand the real-world robustness of existing unsafe image detectors in-depth, we conducted an experiment regarding state-of-the-art, commercially-available detectors by measuring their capability in detecting adversarial unsafe images. Our main objective was to find out whether these existing detectors can be compromised if adversarial perturbations are used to hide unsafe images by malicious users.

In our study, we selected 5 state-of-the-art existing detectors that have the capability to detect unsafe images, namely, Clarifai [8], Google SafeSearch [6], Amazon Rekognition [4], Microsoft Azure [7], and Yahoo Open NSFW [10]. Due to the ubiquity and effectiveness of these detectors, they can be considered as representative of the technology used to defend against unsafe content in existing online platforms. In our experiment, we considered a dataset containing sexually explicit images [43] and used the class labeled "porn" as unsafe images, since all the existing detectors had the capability to detect such images. The outputs given by these existing detectors were varied. The Clarifai, Yahoo Open NSFW, and Amazon Rekognition systems gave a probability score for the unsafe images as outputs. On the other hand, the output of Microsoft Azure was either a true or false label for such images. Google SafeSearch provided even more labels for unsafe images, with the labels being “unknown”, “very unlikely”, “unlikely”, “possible”, “likely”, and “very likely”. Based on these varying methods of measuring whether or not an image is unsafe, we used the following thresholds to determine if an unsafe image is detected. For Clarifai we used a probability score greater than 0.8, for Yahoo Open NSFW and Amazon Rekognition we used a probability score greater than 0.9. For Microsoft Azure we simply considered the true label, and for Google SafeSearch, we considered “likely” and “very likely” labels as unsafe image. The details about how we chose the thresholds for Clarifai, Yahoo Open NSFW, and Amazaon Rekognition can be found in Appendix A.

The Clarifai, Amazon Rekognition, MS Azure, and Google Cloud Vision model weights and architectures were not publicly available, and only the model outputs were accessible to us. The Yahoo NSFW model was a publicly available, open-source model whose weights were accessible. Because many types of adversarial attacks are gradient-based attacks that require the knowledge of the inner workings of the target models, a substitute model, known as a surrogate model was used to emulate the target model [77]. Ideally, the surrogate model should be as close to the target model as possible to generate the most effective attacks. For the non-publicly available detectors, we chose the ResNet-18 model pre-trained on ImageNet as the surrogate model [62], since these models have been shown in the existing literature to be close to the computer vision models employed by public APIs [26]. Similarly, in order to determine how effective the same attacks were on the open-source model, the exact same surrogate model was used to attack the Yahoo Open NSFW model. We used three adversarial attacks to generate adversarial unsafe images on the surrogate model, Square Attack [17], AutoAttack [27], and an ensemble attack consisting of Square Attack and a Gaussian blur. We chose the Square and Square-GB attacks because these are popular black box attacks mentioned in existing literature on adversarial attacks [17] and we chose AutoAttack since this is a recent, state-of-the-art white box attack [27].

We randomly sampled a set of 50 strongly sexually explicit images to craft adversarial images using the three attacks each, after which they were tested against various existing detectors provided through their public APIs, as well as the open-source model. The results of these experiments are shown in Table 2. Each column of Table 2 shows different existing detector attacked, while each row of Table 2 shows the different attacks launched against these detectors. The numbers shown in Table 2 represent the percentage of the submitted images that were classified correctly as sexually explicit. After experimenting with various parameters of adversarial attacks as well as different attack types (the parametric details of these attacks are provided in Appendix B), we found that Square Attack and AutoAttack were particularly effective at fooling unsafe image classifiers with minimal visual perturbations for human observers. To ensure that the images used in this experiment were not too heavily perturbed, an experiment was conducted in which each image was examined to verify that it was identifiable as sexually explicit by four of the authors of this work. In this experiment, images were independently reviewed by four authors and then used majority voting to determine if each image is sexually explicit. All authors were

in unanimous agreement on all images that the images were still sexually explicit despite the presence of adversarial noise. This indicates that the images were visually identifiable as sexually explicit even after perturbations.

From Table 2, we can see that the state-of-the-art unsafe image detectors are indeed successful at detecting non-perturbed sexually explicit images (Table 2, "No Attack"), with an average detection rate higher than 90%. However, they are severely vulnerable to adversarial unsafe images (Table 2, "Square", "Square + GB" and "AutoAttack"). Based on our experiment, adversarial unsafe images crafted using just off-the-shelf adversarial attack algorithms can quite severely compromise existing systems. For instance, the AutoAttack was quite effective at fooling the Amazon Rekognition, MS Azure and Google Safe Search API, with these detectors seeing a 6% or less drop in performance when compared to non-attacked images. Furthermore, Google Safe Search showed a 14% drop in detection accuracy on Square attacked adversarial unsafe images, while the Yahoo Open NSFW model showed a 78% drop in detection accuracy. Our experiment indicates that there is a large security gap between adversarial unsafe images and existing unsafe image detectors, and that this must be immediately addressed. We hypothesize that most adversarial perturbations are located in the high-frequency component of images, however, few works exist that *remove* the perturbation from this perspective. It is critical that we develop techniques to clean the adversarial perturbation in the image for explainability algorithms to correctly identify the unsafe features in an adversarial unsafe image.

# 5 Data

In our work, we considered three unsafe image datasets [2] to demonstrate our system: A sexually explicit images dataset [43], a cyberbullying images dataset [76], and a novel self-harm images dataset.

## 5.1 Sexually Explicit Images

We sampled a subset of the publicly available images dataset [43] for sexually explicit images. This dataset contains 334,327 images from five classes including "neutral", "drawing", "hentai", "porn", and "sexy". We considered "neutral" and "sexy" classes into a single class of non-sexually explicit (i.e., safe) images, and considered the "porn" class for the sexually explicit (i.e. unsafe) images. We considered only the "porn" images for the sexually explicit class of images because they depicted direct sexually explicit content (such as nudity), and "neutral" and "sexy" images as non-sexually explicit because they did not depict any direct sexually explicit content.

[2] We considered three datasets in our work to represent the effectiveness of our system. However, our system is also compatible with other unsafe image categories.

## 5.2 Cyberbullying Images

To analyze the performance of our system on cyberbullying images, we used the dataset introduced by [76], which contains nearly 20,000 images, and are divided into cyberbullying and non-cyberbullying categories. Their cyberbullying dataset was collected from multiple search engines such as Google, Bing, and Baidu, as well as from OSNs such as Instagram, Flickr, and Facebook. From this dataset we perform our experiments on the 5224 cyberbullying samples and 14,628 non-cyberbullying samples. From our observations, the cyberbullying images included content such as rude hand gestures, threatening objects and weapons, or racist or hateful symbols.

## 5.3 Self-harm Images

To demonstrate the capability of our system on the emergent societal issue of self-harm image sharing on OSNs [80], we collected a novel dataset of self-harm images. Our data collection task was approved by our IRB. We collected our dataset by scraping images associated with specific self-harm related tags on Tumblr [9], a popular OSN. To ensure comprehensive coverage, we adopted an incremental approach for collecting tags from Tumblr. Initially, we started the search with the "self harm" tag and then expanded our search by including related tags from the self-harm images that were collected. We iteratively repeated this process until our expanded tag list no longer found new self-harm images. We used the following tags to collect our dataset: *selfharm*, *selfh@rm*, *self h@rm*, *selfmutilation*, *self harm*, *cvtting*, *selfhate*, *s3lfharm*, *self bruising*, *selfbruising*, *tw bruising*, *twcvts*, *selfharn*, *tw cvtting*, *tws3lfharmmcvtting*, *made of styrofoam*, *s3lfharmm*, *tw self hate*, *tw s3lf harm*, *slef harm*, *self mutalition*, *s3lfh4rm*, *cvtt1ng*, *sh*, *tw sh*, *self destruction*, *tw*, *selfhate*, and *shtumblr*.

After sufficient number of images were collected, all images were tested through Google Safe-Search API which has detection capabilities of medical images and violent images. The API returns "unknown", "very_unlikely", "unlikely", "possible", "likely", or "very_likely" tag depending on likelihood of image fitting into the classification, and any image that returns "possible", "likely", or "very_likely" response for "medical" or "violence" tag were collected. Subsequently, the collected images were manually annotated as self-harm and not-self-harm image by visual inspection. Images that contain self-cutting, self-bruising, or anorexia and eating disorder and depicted or promoted self-harm in these ways were annotated as self-harm, while others were annotated as not-self-harm.

After our annotation process, we were left with a dataset of 2,100 self-harm images, which we termed TumblrSelfHarm dataset. Based on a qualitative analysis of our dataset, we observed that the images depicted cutting, bruising, burning, eating disorder behaviors, aftermath of self-harm events such as bloodied bandages, sinks and razors, drawings of self-harm, and images which may be considered to encourage

self-harm and suicide.

# 6 Approach

## 6.1 uGuard Overview

We present the system as data flows through it. The objective of designing our system is a targeted obfuscation of adversarial unsafe images with minimum information loss. This relies on two elements: reconstructing the image so that it is free from adversarial perturbation, and then the use of explainability to target the region to obfuscate. The overview of our system, uGuard is presented in Figure 2. Our system, can be considered in two steps: (1) Building of the adversarially trained robust classifier, and building of the image reconstruction component, and (2) Deployment of the system. To begin building the system, uGuard first takes in datasets of unsafe content. Based on each dataset, our system trains a robust unsafe image classifier model and builds the image reconstruction system. After building these two components for each dataset, the system is ready to deploy. In the deployment stage, uGuard takes in an image which may be adversarially perturbed. Several reconstructed versions of this image are created based on the reconstruction system for each dataset. These reconstructed images are then sent to their respective classifiers. The input image is approved if the reconstructed image is not detected as unsafe. If the reconstructed image is detected as an unsafe image, the image is obfuscated based on the explainability based obfuscation subsystem. If unsafe content is no longer detected by the robust unsafe image detection system after the obfuscation, the obfuscated image is approved. Supposing that the obfuscated image still contains unsafe content, the image will not be approved.

## 6.2 Building of the System

The upcoming subsections describes first the building of the robust classifier, second, the construction of the image reconstruction component, and finally the explainability method for targeted content obfuscation.

### 6.2.1 Robust Classifier to Generate Semantic Features

Deep learning methods, more specifically classification models trained in a supervised fashion aim to generate a model $f \in \mathcal{F}$, such that:

$$\mathbb{E}_{(x,y)\sim\mathcal{P}}[l(f(x),y)] \leq \mathbb{E}_{(x,y)\sim\mathcal{P}}[l(f'(x),y)] \; \forall \; f' \in \mathcal{F},$$

in which the loss function, $l(f(x),y)$, calculates the distance between the predictions of $f(x)$ and what the true label $y$ indicates. A priori the data distribution $\mathcal{P}$ is not know, and we use a training dataset $\mathcal{D}_{tr}$ in an optimization framework to generate the best possible estimator $f$ for the labels observed in the data. Estimating $f$ from the training dataset is known as *empirical risk minimization* (ERM), defined as:

$$\min_{\theta} \quad \sum_{i\in\mathcal{D}_{tr}} l(f(x_i;\theta),y_i) + \lambda\rho(\theta), \tag{1}$$

in which $\theta$ defines the model parameters and $\rho(\theta)$ is a regularization function to constrain the changes of the model parameters at each learning step. We refer to Equation 1 as the baseline model training. We train a Residual Network (RN) for the classification task. We refer to the baseline model $f_{bsl}$, a model trained without any adversarial training or robustness technique (except standard augmentation techniques), such as: Batch Normalization, Dropout, and Parameter Regularization. Any baseline model can be used for the purposes of reconstruction, but it is required that $f_{bsl}$ accurately models the distribution of $\mathcal{D}_{tr}$, and consequently achieves high evaluation accuracy on $\mathcal{D}_{te}$. The high accuracy in the test set, implicates in a good class separation in the feature space, and consequently very distinct distributions between the classes. To train the our robust model, we initially construct a set $\mathcal{R} = \{R(x_i), x_i, y_i\}$ of the latent representations extracted from dataset $\mathcal{D}_{tr}$ by model $f_{bsl}$, and its originating images-labels pair. The latent representations $R(x_i)$ correspond to the set of features the model extracts just after all the convolutional layers, and just before the set of fully connected layers. We use $f_\theta(.)$ to generate the set $\mathcal{R}$ that contains the latent representations for all samples of all classes in $\mathcal{D}_{tr}$.

We generate one set of latent representations $\Psi_j$ for each class. For each $\Psi_j$ we obtain the mean $\mu_{\Psi_j} \in \mathbb{R}^k$ as the average of each individual component of each $R(x_i) \in \Psi_j$, and the covariance:

$$\sigma_{\Psi_j} = \mathbb{E}[(\Psi_j - \mathbb{E}[\Psi_j])(\Psi_j - \mathbb{E}[\Psi_j])^T] \tag{2}$$

where T is the transpose operator.

Each $\Psi_j$ represents a set of semantic features of each class. These semantic features are translated from the originating images $x_i$. These training images are the base to create feature dictionaries, which are the base of our reconstruction algorithm. We generate reconstruction dictionaries using *Convolutional Dictionary Learning* (CDL). Specifically, given a set of images $x_i \in \Psi_j$ composed of $S$ training images $\{x_t\}_{s=1}^{S}$, CDL is implemented through minimizing:

$$\begin{aligned} \min_{\{d_m\},\{r_{s,m}\}} \quad & \frac{1}{2}\sum_{1}^{S}\left\|\sum_{1}^{M} d_m * r_{s,m} - x_s\right\|_2^2 \\ & + \lambda\sum_{1}^{S}\sum_{1}^{M}\|r_{s,m}\|_1 \\ \text{s.t.} \quad & \|d_m\|_2 \leq 1, \forall m \in 1,...,M \end{aligned} \tag{3}$$

where $d_m$ are the $M$ atoms that comprise the dictionary $\Omega$, and $r_{s,m}$ are a set of coefficient maps, defined as:

$$\min_{\{r_m\}} \quad \frac{1}{2}\left\|\sum_{1}^{M} d_m * r_m - x\right\|_2^2 + \lambda\sum_{1}^{M}\|r_m\|_1 \tag{4}$$

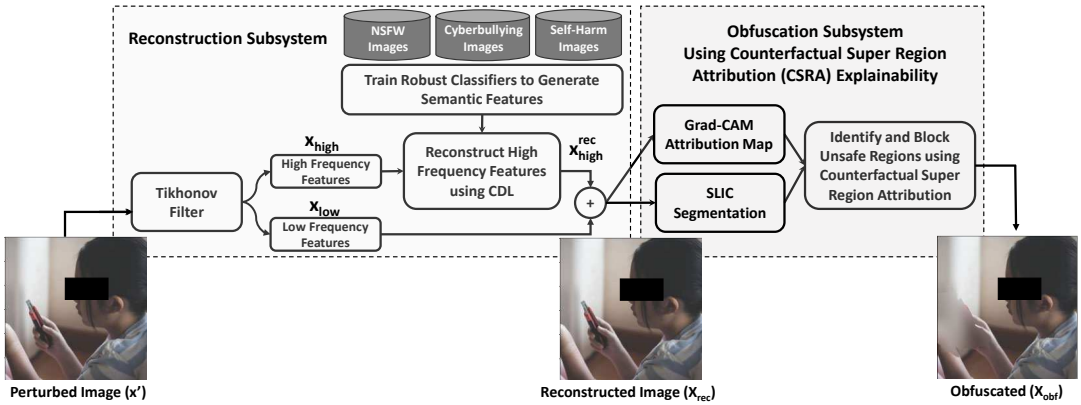


Figure 2: Overview of UGUARD.

CDL is a computationally expensive algorithm that does not scale well to larger images and datasets. We use the optimized version proposed by [51], to minimize the convergence time and ADMM [21] to solve the minimization problem. The images, class distributions, and class reconstruction dictionaries generated for all classes are utilized for the semantic reconstruction dictionary, $\Phi = \{D, \Psi, (\mu_\Psi, \sigma_\Psi)\}$.

The model is capable of generalizing to any input which falls within the same distribution as the train and test set. But such assumption does not account for adversarial inputs. In fact, (Equation 1) is highly vulnerable to small perturbations, crafted by adversarial algorithms. In a general formulation, these perturbations are generated such that:

$$\max_{\|\delta\|_2 \leq \varepsilon} \quad l(f(x_i + \delta; \theta), y_i). \tag{5}$$

By introducing the perturbation $\delta$ in $\mathcal{D}_{te}$, we shift the actual test distribution to the tail of the training distribution, effecting the performance of $f(x)$ when evaluated in $\mathcal{D}_{te}$. Curently, the most used technique to deal with these adversarial attacks is adversarial training, defined as:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\delta\in\Delta} l(f(x+\delta), y) + \lambda\rho(\theta), \tag{6}$$

Equation 6 is the *standard adversarial training* and addresses the immediate issue of adversarial samples crafted to attack $f(.)$. While the adversarial attack strategy of a min-max optimization shown in Equation 6 has shown very successful results, it fails in generalizing the method to unseen attacks [70]. This occurs because the network does not learn to extract robust latent representations, but rather learns to change the FC layers to classify latent representations extracted from adversarial and clean samples in the same class. To address this issue we change the standard adversarial training equation, adding an extra constraint in the objective function:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \max_{\delta\leq\varepsilon} l(f(x'), y) + \lambda \|\theta)\|_2^2 + \alpha(R(x') - \mu)\sigma^{-1}(R(x') - \mu))^{\frac{1}{2}} \tag{7}$$

where the last term, the *Mahalanobis Distance* (MD), minimizes the distance between the extracted adversarial latent representations $R(x')$ and the cluster distribution, following the association in $\Phi$. Minimizing the distance between the currently extracted features and the clean model feature distribution, allows the model to learn to extract meaningful features, rather than learning the specific adversarial attack pattern present in the training set.The robust semantic model is defined $f_{rob}$, and latent representations extracted from input, $x_i$, with $f_{rob}$ as $R_{rob}(x)$.

### 6.2.2 Image Reconstruction

We assume all input to our system is potentially adversarial. Previous works have shown that high-frequency signals play an important role in the generation of adversarial images [86]. Adversarial images undergo a transformation such that the feature activations are similar to those of the target. As a consequence based on the class defined by $f(x'_i)$, we select the semantic reconstruction dictionary which best reconstructs the high-frequency components of $x'_i$. We use the dictionary which the feature distribution minimizes the $MD(\Phi, R_{rob}(x'_i))$. In parallel, we decompose $x'_i$ into a high-frequency component, $x'_{high}$, and a low frequency component, $x'_{low}$, using the Tikhonov filter [36]:

$$\arg\min_{x_{low}} \quad \frac{1}{2}\|x_{low} - x\|_2^2 + \frac{\lambda}{2}\sum_j \left\|G_j x_{low}\right\|_2^2$$

where $G_j$ is an operator that computes the discrete gradient along image axis $j$. Therefore, $x'_{high} = x' - x'_{low}$.

The reconstruction of $x'_{high}$ follows the standard sparse coding representation:

$$x_{high}^{rec} \approx Dr = d_1 r_1 + \cdots + d_M r_M,$$

in which $D$ is the dictionary learned only from patches of clean images. This leads to a high-frequency component constituted of only class specific features learned from the clean images, and free of manipulation.

**Algorithm 1** CSRA Algorithm

$NumRegions = m$
$Compactness = n$
$NumROI = k$
**Input:** Model, Image
$AttrMap = CAMAttrScores(Model, Image)$
$Seg = SLIC(Image, NumRegions, Compactness)$
$AvgAttrs = \{\}$
**for** $S \in Seg = \{S_1, S_2, ..., S_{NumRegions}\}$ **do**
    $NumPixInSeg = |S_i|$
    $AvgRegionAttr = \frac{\sum_{n=1}^{NumPixInSeg} S_{in} \in AttrMap}{NumPixInSeg}$
    $AvgAttrs.append(AvgRegionAttr)$
**end for**
$TopAttrs = argmax_{AvgAttrs' \subset AvgAttrs, |AvgAttrs'|=NumROI} \sum_{a \in AvgAttrs'} a$
$Powerset = \mathcal{P}(TopAttrs)$
$ImgVers = MaskAttrs(Image, Powerset)$
$Scores = \{\}$
**for** $i \in ImgVers = \{i_1, i_2, ..., i_{2^{NumROI}}\}$ **do**
    $Score = Softmax(Model, i) + \frac{(NumMaskedPixels)}{(NumOfPixels)}$
    **if** ModelPred is Class of Interest **then**
        $Score$ += 1
    **end if**
**end for**
$TopImageVersion \leftarrow LowestScoreImgInImgVers$
**Output:** $TopImageVersion$

The final image is obtained by adding the low and high-frequency components of the image:

$$x_{rec} = x_{low} + x_{high}^{rec} \tag{8}$$

#### 6.2.3 Explainability-based Content Obfuscation

We define two objectives for the obfuscation of unsafe image content. The first objective is to obfuscate the unsafe content in the unsafe image such that is no longer unsafe. The second objective is to minimize the region that is obfuscated to just the unsafe region. We found that existing explainability techniques faced challenges to meet both of these objectives. The explainability-based obfuscation method which we call Counterfactual Super Region Attribution (CSRA) that we used is detailed in Algorithm 1. Our method combines information from grayscale attribution maps output by Grad-CAM methods with features generated by SLIC superpixel segmentation [12].

SLIC superpixel segmentation is a K-means clustering based method in the 5-D space of RGB color and x,y pixel coordinates [12], and is an effective superpixel segmentation method. SLIC requires two parameters: (1) number of superpixel regions, and (2) compactness. Compactness defines the clustering's focus between color information and pixel location for the generation of the superpixels, where higher compactness leads the clustering to have more emphasis on pixel location. High compactness values will lead to more box-like superpixels.

Unlike other perturbation-based methods, we leverage gradient-based attribution maps like Grad-CAM [68] to lead an informed approach to sampling combinations of superpixel regions. We hypothesize that by limiting ourselves to just the regions of the image that the attribution map deems important, we can avoid sampling regions that are unlikely to have any contribution to the model decision. CSRA takes in an integer value of *NumROI*, which is the number of regions of interest. We only consider superpixels that are in the top *NumROI* of highest average contribution based on the attribution map. Next, we create a Power Set of those high attribution superpixels that were identified. Next, we take that Power Set of superpixels and use them to create different versions of the input image by replacing the superpixels in the set with black pixels. From there, we evaluate each version of the image based on the softmax score output by the model on each version of the image, and the proportion of the image that has been replaced by black pixels. Based on this evaluation, we output the version of the image that has the lowest score according to a score function that penalizes versions of the image that are detected as unsafe, and that penalizes obfuscation.

Compared to other perturbation-based methods, our CSRA approach allows us to perform a more thorough analysis of which superpixels are truly important to the model's decision. Because of this, we are able to identify the regions of the image that cause the unsafe image to be unsafe, and then obfuscate them, while minimizing the total area that is obfuscated.

By just limiting to the *NumROI*, we may not be able to identify all of the unsafe regions in the image. To manage cases where the initial obfuscation of the regions identified still results in an unsafe image, the CSRA algorithm has an additional safeguarding method. We call this additional safeguarding method Limited Region Dilation. With Limited Region Dilation we iteratively expand the regions identified by CSRA by 5 pixels until either unsafe content is no longer detected or until some percentage threshold of the image has been obfuscated.

## 7 System Implementation and Evaluation

### 7.1 Implementation

For the standard classifier that we compare our UGUARD's unsafe image detection capabilities to, we trained a ResNet-50 classifier using Pytorch [62] libraries using pre-trained model weights trained from the ImageNet dataset [30]. The sexually explicit, cyberbullying and self-harm datasets were each divided into train, validation, and test sets, with 80% being allocated to the train set, and 10% each allocated to validation

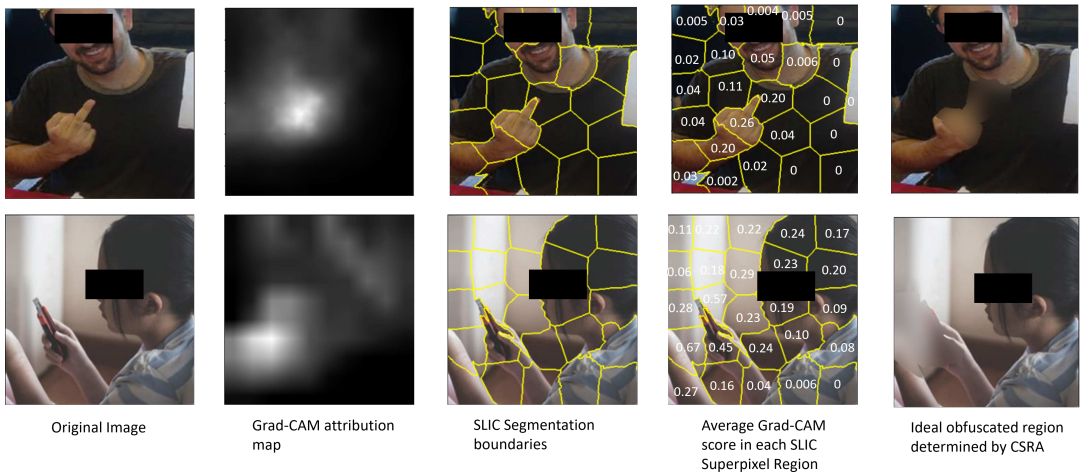


Figure 3: Examples from cyberbullying and self-harm content categories showing the steps in the CSRA algorithm.

and test sets. We trained our models for 50 epochs and saved the trained weights of the models that have the highest classification accuracy on the validation set. All of our evaluations are performed on the test set. We take the same steps and use the same model architecture to train our robust models, but adversarially perturb a proportion of the datasets before training. For the implementation of the CSRA explainability method, we use the Pytorch Grad-CAM library by [37] to get the Grad-CAM maps, and we use the scikit-image library to generate features using SLIC superpixel segmentation.

## 7.2 Evaluation

We summarize the evaluation of our system as follows: (i) Evaluation of the effectiveness of UGUARD's reconstruction and robust model components on adversarial unsafe images, showcasing UGUARD's resistance to such images (Section 7.3). (ii) Effectiveness of UGUARD's explainability-based content obfuscation component (CSRA), showing it's ability to make unsafe images safe through content obfuscation, while minimizing the safe information lost (Section 7.4). (iii) A study that evaluates the number of regions considered for obfuscation by the CSRA algorithm to make unsafe images safe for different datasets (Section 7.4). (iv) An end-to-end evaluation of the UGUARD system on how it handles adversarial unsafe content versus state-of-the-art systems. We show detection accuracy, the success of making unsafe images safe, and the amount of the image that is obfuscated using our system (Section 7.5). (vi) A user experiment with online participants indicating that OSN users prefer the regional obfuscation approach employed by UGUARD to the full obfuscation approach used by many OSNs (Appendix C).

## 7.3 Effectiveness of Image Reconstruction in UGUARD Against Adversarial Unsafe Images

We evaluate the effectiveness of our image reconstruction and robust model component on two of the attacks that the robust model was trained against and on two attacks the model was not trained against to demonstrate robustness against unknown attackers. In the dataset to train the robust model, we include the PGD [52], BIM [48] and Square attack [17] attacks, implemented with the Torchattacks [44] library. For PGD we set the following parameters: $\varepsilon = 8/255$, $\alpha = 2/255$, and steps = 10. For BIM, we set $\varepsilon = 4/255$, $\alpha = 1/255$, and steps = 10. Finally, for Square attack, we set n_queries = 500, n_restarts = 1, and $\varepsilon = 16/255$. For the training of the robust model, we evenly divide the training data into four parts to include equal representations of unattacked data, PGD, BIM and Square attacked data. To represent attacks unknown to our robust model, we test against images perturbed by AutoAttack [27] and DeepFool [54]. The results of the robustness experiments are shown in Table 3. For unattacked images, we see that for sexually explicit and self-harm images there is a slight drop in accuracy when comparing the baseline model to our robust system. A small drop in classification accuracy on unattacked data is a common observation when comparing non-robust versus robust models [82]. However, this drop in accuracy is more than made up for when comparing the two models on attacked data. For instance, the BIM attack on the sexually explicit baseline model showed an accuracy of 25.17% whereas the UGUARD's sexually explicit detector had an accuracy of 88.74% on BIM attacked data. Across the three unsafe image detectors in the UGUARD system, we see very small drops in classification accuracy from unattacked data to these attacked data, with an average drop in accuracy of just 1.32%. Our experiments show that AutoAttack and DeepFool were particularly strong attacks against the baseline model. However, despite our model having not been trained against these attacks, we show that UGUARD has minimal drops in accuracy. Additionally, we test on images perturbed by BIM and DeepFool to show that UGUARD can be robust against a combination of attacks. These experiments show that the robust model in conjunction with the image reconstruction processing step demonstrates robustness against multiple seen

| | | Known Attacks | | Unknown Attacks | | |
|---|---|---|---|---|---|---|
| | No Attack | BIM | Square | AutoAttack | DeepFool | BIM+DF |
| Baseline | - | - | - | - | - | - |
| ResNet-50 | - | - | - | - | - | - |
| Sexually Explicit | 93.33 | 25.17 | 63.58 | 0 | 12.58 | 59.60 |
| Cyberbullying | 91.14 | 48.34 | 56.85 | 1.87 | 5.23 | 69.54 |
| Self-Harm | 94.70 | 53.64 | 53.64 | 5.23 | 5.23 | 72.18 |
| UGUARD | - | - | - | - | - | - |
| Sexually Explicit | 90.06 | 88.74 | 88.07 | 84.11 | 88.07 | 89.40 |
| Cyberbullying | 96.02 | 94.70 | 95.36 | 90.73 | 92.71 | 93.38 |
| Self-Harm | 90.73 | 88.74 | 90.07 | 86.75 | 90.73 | 90.07 |

Table 3: Comparison of baseline classifiers vs. UGUARD.

and unseen attacks. In Table 3 we compare the accuracy of each unsafe image category for unattacked data, BIM, Square, AutoAttack, and DeepFool attacked images between a standard ResNet-50 classifier, and the UGUARD system. It can be seen that the standard ResNet-50 models show significant performance drops against these attacks, whereas the UGUARD system sees minimal drops in accuracy on these attacks. This indicates that UGUARD is suitable for detecting adversarial unsafe images.

While our experiments show good robustness against unseen attacks, due to the ever evolving landscape of attacks, continued robustness against unseen attacks is uncertain. Once a new attack is identified to be successful in fooling our system, our system can be easily updated. It only requires that we include the new attack in the dataset of attacked images that we train the robust model on. By utilizing the previous robust model's weights, the model can be trained quickly and made robust against the new attack.

## 7.4 Effectiveness and Impact of CSRA Explainability on Targeted Obfuscation

**Effectiveness.** We conduct a perturbation analysis [84] to evaluate the effectiveness of our CSRA algorithm for unsafe content obfuscation based on explainability. The perturbation analysis tests how the model predictions change when perturbing specific regions that are identified by the explainability method. This tests the ability of the explainability method to localize the features of interest. We follow the strategy of Fu et al., who masks the pixels corresponding to the top 20% of the values in the attribution map in order to evaluate CAM based methods [35].

We evaluate our method, CSRA, against multiple Grad-CAM based methods (Grad-CAM [68], XGrad-CAM [35], Grad-CAM ++ [24], FullGrad [71]) and LIME [63], a perturbation based method. Our evaluation compares the image classifier explanation methods on their ability to change the classifier decision, and on the proportion of the image that was masked. In order to evaluate the ability to change the classifier decision, we take the softmax probability output by the model and use the class with the highest probability to be the model's decision. To test CSRA's ability to change the classifier's decision, we simply take the image output by CSRA and run it through the classifier. For our experiments on CSRA we define the number of superpixel regions to be 30 and compactness to be 50. For Grad-CAM methods, we identify the pixels that are in the top 20% of the values in the attribution map and replace them with black pixels in the original image, and then send it to the classifier. For LIME, we take the regions identified as contributing to the model's decision and mask them with black pixels. We set the number of samples that LIME uses to learn the linear model that is used to generate the explanations equal to 256. Each of these samples have different regions of the image perturbed in order to learn the regions that contribute to the model's decision. We choose 256 as the number of samples for LIME because for CSRA with *NumROI* = *8*, CSRA also analyzes 256 different perturbed samples. This allows us to compare these two methods on similar grounds of computational power required. When combining CSRA with Limited Region Dilation, we set the threshold for Limited Region Dilation to be 50%.

We test on 151 samples from each dataset that have an unsafe ground truth label and are predicted as unsafe by our classifiers. The results of these experiments are displayed in Table 4. Across the three unsafe image datasets, we found that Grad-CAM tended to change more model decisions than the derivative methods of Grad-CAM. Our experiments show that when combined with Limited Region Dilation, CSRA is able to effectively obfuscate the unsafe regions of the unsafe images while minimizing the amount of obfuscation. For CSRA with *NumROI* = *8*, we show that CSRA outperforms Grad-CAM based methods across all of the different unsafe image detectors. The experiment shows that LIME is able to successfully obfuscate the unsafe regions of unsafe images, however, this results in excessive amounts of obfuscation, with over 60% of the image being obfuscated on average. Overall, the results show that CSRA is able to significantly outperform popular gradient based and perturbation based methods in their ability to localize the regions of the image that are unsafe.

**Impact Analysis.** We introduced CSRA for our unsafe image obfuscation subsystem, based on explainability. In Figure 4 we show how the CSRA image explanation method performs as a method for detecting the most important parts of an image at different *NumROI*. As expected, increasing the *NumROI* increases the probability that CSRA has identified the most important parts of the image. Our proposed content moderation solution combines CSRA with Limited Region Dilation, but as shown in Figure 4, CSRA by itself can successfully identify the most unsafe parts of the image at large enough values of *NumROI*. For instance, on cyberbullying images, with *NumROI* = 8, we are able to make 94.67% of cyberbullying images safer, while only obfuscating 12.73% of the overall image.

The results from Figure 4 show that the level of obfuscation necessary to make an unsafe image safer varies between datasets. In particular, the explanations on the sexually explicit dataset and model resulted in identifying a greater proportion of the image when compared to the cyberbullying and self-harm experiments. From qualitative observations of ob-

| | CSRA | | CSRA + LRD | | Grad-CAM | | XGrad-CAM | | Grad-CAM ++ | | FullGrad | | LIME | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. | % of Pred. Changed | % of Image Obf. |
| Sexually Explicit | 70 | 21.69 | 96.67 | 27.00 | 48 | 20 | 48 | 20 | 42.67 | 20 | 26.67 | 20 | 100 | 66.19 |
| Cyberbullying | 94.67 | 12.73 | 99.5 | 13.37 | 81.33 | 20 | 73.33 | 20 | 60 | 20 | 84 | 20 | 100 | 60.85 |
| Self-Harm | 86.67 | 10.79 | 94.67 | 14.00 | 62.67 | 20 | 58 | 20 | 25.33 | 20 | 68.67 | 20 | 100 | 76.90 |

Table 4: Quantitative results of different explainability methods for content obfuscation.

fuscated images in our experiments, we found that the CSRA algorithm was often identifying nudity as unsafe. In sexually explicit images, the nude body often takes up a large part of the total image, which causes the model to only deem the image as a safe image after covering more of the body. On the other hand, the unsafe regions of self-harm and cyberbullying images often make up a smaller portion of the image. In the cyberbullying images, the unsafe portion of the image is made up of rude hand gestures, or the brandishing of a weapon. For self-harm images, the unsafe region is often the blade near the skin or cutting wounds. In these cases, the unsafe part of the image makes up less of the total area of the image.

## 7.5 End-to-end Evaluation of uGuard

uGuard is an end-to-end system that takes in adversarial unsafe images and outputs safer images for viewers. To evaluate how well uGuard works end-to-end to transform adversarial unsafe images into safer images, we evaluate on samples that have been perturbed with Square attack, which were shown to be very effective on public unsafe image detection systems. In Table 5 we average the detection performance of the 5 existing state-of-the-art detectors that we tested in Section 4 and compare it to uGuard's detection performance. We show that uGuard detects these perturbed samples with an average accuracy of 91.67% across sexually explicit, cyberbullying, and self-harm datasets, compared to the average sexually explicit detection accuracy of 45.60% on the existing detectors. Furthermore, Table 5 shows that uGuard's explainability based obfuscation method makes 96.94% of these unsafe samples safer, with an average obfuscation of just 18.03%. We show examples of how our CSRA-based obfuscation method manages cyberbullying and self-harm samples in Figure 3. This figure shows how the CSRA method considers the superpixels with the highest average Grad-CAM scores, and then obfuscates based on the superpixels that are most important according to CSRA's counterfactual analysis, which minimizes the obfuscation to just the unsafe part of the image. Consequently from an end-to-end perspective, uGuard is successful in managing adversarial unsafe content.

In order to gain some insight into the obfuscation preferences of social media users, we conducted a survey of 100 Amazon MTurk workers who were asked to evaluate a potentially sensitive image obfuscated with regional obfuscation versus obfuscation of the whole image. Our study shows that, on average, users rate the partially obfuscated image as providing more information, more satisfying, and as having a greater sense of human contact than the fully obfuscated image. Further details about our MTurk social media user survey can be found in Appendix C.

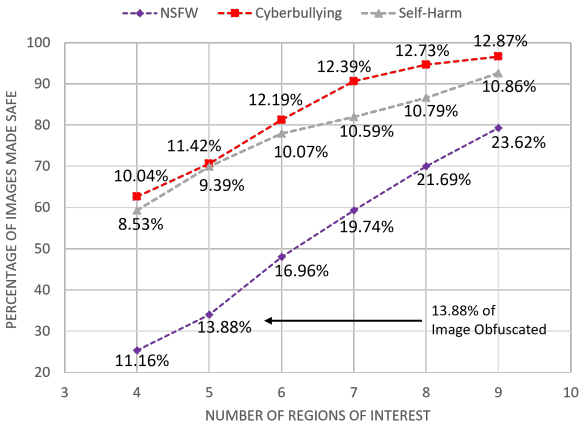


Figure 4: Impact of *NumROI* on CSRA explanations quality.

## 7.6 Running uGuard on Images in the Wild

To test the generalizability of our approach, we evaluated our system on 1020 images from three sources of unsafe images in the wild (depicted in Table 6). 4chan is well known for spreading celebrity nude images from an event known as "Celebgate" [1]. While the object of discussion on this thread was about a specific celebrity, many such cases of deepfake porn about specific individuals are known to exist. To evaluate uGuard on sexually explicit images, we found an online discussion board on 4chan [11] that focuses on the sharing of adult content. We captured 510 instances of sexually explicit images from the top 100 threads on this discussion board and ran them through our system. We conducted our obfuscation experiment with CSRA *NumROI* = *12* using 510 images that were both labeled as sexually explicit and were detected as sexually explicit. In this study, the resulting obfuscated

| | Public API | uGuard | | |
|---|---|---|---|---|
| | Adversarially Perturbed Accuracy % | Adversarially Perturbed Accuracy % | % Adversarially Perturbed Images Obf. to be Safer | Obfuscation % |
| Sexually Explicit | 45.60 | 88.07 | 96.67 | 27.00 |
| Cyberbullying | N/A | 95.36 | 99.50 | 13.37 |
| Self-Harm | N/A | 90.07 | 94.67 | 14.00 |

Table 5: Management of adversarial unsafe images.

| | Source | Source Description | No. of Successfully Obfuscated Unsafe Images |
|---|---|---|---|
| Sexually Explicit | 4chan [11] | 4chan board for sharing adult GIFs. | 445 out of 510 |
| Self-Harm | BestGore [5], Twitter | #self-harm tagged images. | 488 out of 510 |

Table 6: Running uGUARD on images in the wild.

images were independently visually inspected by the authors and deemed successful if the authors unanimously perceived them to be safe. 445 out of 510 sexually explicit images were successfully obfuscated by our system. A member of a federal agency (who is also a collaborator in this work) directed our attention to an archive of the site BestGore [5]. We collected 510 self-harm images from Twitter and BestGore [5], searching for "selfharm" tagged content. We conducted our obfuscation experiment with CSRA *NumROI = 12* using 510 images that were both labeled as self-harm and were detected as self-harm. 488 out of 510 self-harm images were successfully obfuscated by our system. In total, 91.47% of these unsafe images were made safer.

# 8 Discussion

## 8.1 Limitations

We discuss some potential limitations of our work. A potential limitation of our work is that the datasets used in our experiments may not be representative of people with different skin tones and gender expressions. A more comprehensive study of this should be performed in future work. We only collected publicly available images, and we were unable to collect self-harm images from private posts or posts that were extremely sensitive and hence not available in public domain. As a result, our dataset may not be fully representative of this problem. Secondly, despite our testing of the reconstruction component on adversarial attacks that uGUARD was not trained on, our system may still be vulnerable to adversarial attacks. However, a study into the vulnerabilities of image reconstruction is beyond the scope of this work, and will be investigated in future work. Therefore, we can only claim that uGUARD is suitable for known adversarial attacks. Lastly, uGUARD is applicable to unsafe images where specific regions are the causes of the image to be unsafe, e.g., the genitalia regions in sexually explicit images. However, uGUARD may not be applicable to unsafe images where the unsafe region is not based on distinct visual regions, such as hateful memes, where the unsafe region is a combination of the image content and text overlaid on the image, or screenshots of hateful text.

## 8.2 Ethical Considerations

Our data collection task and user study were approved by our institution's IRB. Our IRB protocol put forth several ethical standards pertaining to crucial aspects of our research, including sensitive data handling, participant consent and researcher's well-being that our team strictly monitored and followed throughout the course of this work to ensure the safety of not only the subjects depicted in the images in our dataset and the participants surveyed but also the researchers in our team who were involved in these processes. We have included a few image samples in this paper to help readers better understand our paper while taking steps to ensure no harm to the reader as well as the people pictured by masking their identities and other sensitive parts. Furthermore, we have followed standard ethical guidelines when analyzing the data and presenting the results, including safely storing data, protecting the anonymity/privacy of the users, and not attempting to track users across websites [64].

## 8.3 Extending uGUARD to Other Unsafe Content

Our system can be conveniently extended with new unsafe content categories by integrating the dataset used for the new content. For example, uGUARD can be extended to include Non Consensual Intimate Imagery (NCII) or Child Sexual Abuse Material (CSAM), by adding datasets of clean images of these categories and their classifiers to the Reconstruction Subsystem. This extendibility also allows uGUARD to be flexible to accommodate the cultural norms of the country or region that uGUARD is deployed in. The cultural norms of the society can dictate what is considered as unsafe image content. While the depictions of certain forms content may be deemed improper or even illegal in some countries, in other countries this content might not cause any concerns. Furthermore, our system allows for flexibility in the amount of obfuscation that can be applied in the automated moderation process by changing the threshold of obfuscation in Limited Region Dilation.

# 9 Conclusion and Future Work

In this paper we investigate unsafe image detection systems and automated content moderation of unsafe images. We shown that state-of-the-art systems that detect unsafe image content are vulnerable to adversarial unsafe images, and that existing explainability techniques are not suitable for automated content obfuscation. To solve these deficiencies, we introduce our system, uGUARD. Our evaluations shows that uGUARD is highly effective in neutralizing the threat of adversarial unsafe images. As part of our future work, we plan to include other unsafe content categories into our system. For example, another category of content that social media companies have policies against is around extreme violence and gore.


# Acknowledgments

This research project and the preparation of this publication were funded in part by the Department of Home-

land Security (DHS), United States Secret Service, National Computer Forensics Institute (NCFI) via contract number 70US0920D70090004 and by NSF Grant No. 2245983.

## References

[1] Celebgate. https://www.nbcnews.com/pop-culture/pop-culture-news/almost-600-accounts-breached-celebgate-nude-photo-hack-fbi-says-n372641, 2015.

[2] Judge OKs $85 mln settlement of Facebook moderators' PTSD claims. https://www.reuters.com/legal/transactional/judge-oks-85-mln-settlement-facebook-moderators-ptsd-claims-2021-07-23/, 2021.

[3] Social Media Fact Sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/, 2021.

[4] Amazon rekognition. https://aws.amazon.com/rekognition/, 2022.

[5] Bestgore.fun. https://bestgore.fun, 2022.

[6] Detect explicit content (safesearch). https://cloud.google.com/vision/docs/detecting-safe-search, 2022.

[7] Microsoft azure detect adult content. https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-detecting-adult-content, 2022.

[8] Nsfw model for content detection. https://www.clarifai.com/models/nsfw-model-for-content-detection, 2022.

[9] Tumblr. https://www.tumblr.com/, 2022.

[10] Yahoo open nsfw. https://github.com/yahoo/open_nsfw, 2022.

[11] 4chan. 4chan gif. https://boards.4chan.org/gif/, 2023.

[12] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[13] Rachel A Adler and Spring Chenoa Cooper. "when a tornado hits your life:" exploring cyber sexual abuse survivors' perspectives on recovery. *Journal of Counseling Sexology & Sexual Wellness: Research, Practice, and Education*, 4(1):1–8, 2022.

[14] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560, 2018.

[15] Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics*, 18(8):5031–5042, 2022.

[16] Monica Anderson and Jingjing Jiang. Teens, Social Media and Technology 2018. https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/, 2018.

[17] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501, 2020.

[18] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *USENIX Security Symposium*, 2022.

[19] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. In *ACM Special Interest Group on Security, Audit and Control*, pages 2154–2156, 2018.

[20] Matt Binder. Facebook claims its new ai technology can automatically detect revenge porn. https://mashable.com/article/facebook-ai-tool-revenge-porn, 2019.

[21] Stephen Boyd, Neal Parikh, and Eric Chu. *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc, 2011.

[22] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on lidar-based perception in autonomous driving. In *ACM Special Interest Group on Security, Audit and Control*, pages 2267–2281, 2019.

[23] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

[24] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 839–847, 2018.

[25] Charalampos Chelmis and Mengfan Yao. Minority report: Cyberbullying prediction on instagram. In *ACM Conference on Web Science*, pages 37–45, 2019.

[26] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. In *Conference on Neural Information Processing Systems*, 2019.

[27] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

[28] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, pages 196–204, 2018.

[29] Jelle Demanet, Kristof Dhont, Lies Notebaert, Sven Pattyn, and André Vandierendonck. Pixelating familiar people in the media: Should masking be taken at face value? *Psychologica Belgica*, 47(4):261–276, 2007.

[30] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[31] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[32] Carmen Esposito, Gregory A Landrum, Nadine Schneider, Nikolaus Stiefl, and Sereina Riniker. Ghost: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, 61(6):2623–2640, 2021.

[33] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[34] Facebook. Facebook community standards. https://transparency.fb.com/policies/community-standards/, 2023.

[35] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In *British Machine Vision Conference*, 2020.

[36] Cristina Garcia-Cardona and Brendt Wohlberg. Convolutional dictionary learning: A comparative review and new algorithms. *IEEE Transactions on Computational Imaging*, 4(3):366–381, 2018.

[37] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[39] Trisha Hendricks. Cyberbullying increased 70% during the pandemic; arizona schools are taking action. https://www.12news.com/article/news/crime/cyberbullying-increased-70-during-the-pandemic-arizona-schools-are-taking-action/75-fadf8d2c-cf11-43f0-b074-5de485a3247d, 2021.

[40] Steven Hill, Zhimin Zhou, Lawrence K Saul, and Hovav Shacham. On the (in) effectiveness of mosaicing and blurring as tools for document redaction. In *Privacy Enhancing Technologies*, number 4, pages 403–417, 2016.

[41] Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. Detection of cyberbullying incidents on the instagram social network. *International Journal of Recent Engineering Science*, 2015.

[42] Xin Jin, Yuhui Wang, and Xiaoyang Tan. Pornographic image recognition via weighted multiple instance learning. *IEEE Transactions on Cybernetics*, 49(12):4412–4420, 2018.

[43] Alex Kim. Nsfw data scraper. https://github.com/alex000kim/nsfw_data_scraper, 2021.

[44] Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

[45] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14254–14263, 2020.

[46] Pavel Korshunov, Andrea Melle, Jean-Luc Dugelay, and Touradj Ebrahimi. Framework for objective evaluation of privacy filters. In *Applications of Digital Image Processing*, volume 8856, 2013.

[47] Meredith Krause. Identifying and managing stress in child pornography and child exploitation investigators. *Journal of Police and Criminal Psychology*, 24(1):22–29, 2009.

[48] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall, 2018.

[49] Yifang Li, Nishant Vishwamitra, Hongxin Hu, and Kelly Caine. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[50] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. In *ACM on Human-Computer Interaction*, volume 1, pages 1–24, 2017.

[51] Jialin Liu, Cristina Garcia-Cardona, Brendt Wohlberg, and Wotao Yin. First-and second-order methods for online convolutional dictionary learning. *SIAM Journal on Imaging Sciences*, 11(2):1589–1628, 2018.

[52] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[53] Cade Metz. Twitter's new ai recognizes porn so you don't have to. https://www.wired.com/2015/07/twitters-new-ai-recognizes-porn-dont/, 2015.

[54] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

[55] Alan Mozes. Teen social media posts about cutting, self-harm are soaring. https://www.webmd.com/parenting/news/20211117/teen-social-media-posts-about-cutting-other-self-harm-are-soaring, 2021.

[56] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020.

[57] NCPR. Information and resources to curb the problem of cyberbullying. https://www.ncpc.org/resources/cyberbullying/, 2009.

[58] Virginia Negri, Dario Scuratti, Stefano Agresti, Donya Rooein, Gabriele Scalia, Amudha Ravi Shankar, Jose Luis Fernandez Marquez, Mark James Carman, and Barbara Pernici. Image-based social sensing: combining ai and the crowd to mine policy-adherence indicators from twitter. In *IEEE/ACM International Conference on Software Engineering: Software Engineering in Society*, pages 92–101, 2021.

[59] Quang-Huy Nguyen, Hoang-Loc Tran, Thanh-Thien Nguyen, Dinh-Duy Phan, Duc-Lung Vu, et al. Multi-level detector for pornographic content using cnn models. In *RIVF International Conference on Computing and Communication Technologies*, pages 1–5, 2020.

[60] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *ACM ASIA Conference on Computer and Communications Security*, pages 506–519, 2017.

[61] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. Sok: Security and privacy in machine learning. In *IEEE European Symposium on Security and Privacy*, pages 399–414, 2018.

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems*, pages 8024–8035. 2019.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *ACM Special Interest Group on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[64] Caitlin M Rivers and Bryan L Lewis. Ethical research standards in a world of big data. *F1000Research*, 3(38), 2014.

[65] Laura Sanchez, Cinthya Grajeda, Ibrahim Baggili, and Cory Hall. A practitioner survey exploring the value of forensic tools, ai, filtering, & safer presentation for investigating child sexual abuse material (csam). *Digital Investigation*, 29, 2019.

[66] Sebastian Scherr, Florian Arendt, Thomas Frissen, and José Oramas M. Detecting intentional self-harm on instagram: development, testing, and validation of an automatic image-recognition algorithm to discover

cutting-related posts. *Social Science Computer Review*, 38(6):673–685, 2020.

[67] SchoolSafety.gov. Bullying and cyberbullying. https://www.schoolsafety.gov/bullying-and-cyberbullying, 2019.

[68] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[69] Samuel Henrique Silva, Arun Das, Adel Aladdini, and Peyman Najafirad. Adaptive clustering of robust semantic representations for adversarial image purification on social networks. In *International AAAI Conference on Web and Social Media*, volume 16, pages 968–979, 2022.

[70] Samuel Henrique Silva and Peyman Najafirad. Opportunities and challenges in deep learning adversarial robustness: A survey. *arXiv preprint arXiv:2007.00753*, 2020.

[71] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. In *Conference on Neural Information Processing Systems*, volume 32, 2019.

[72] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[73] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

[74] Tim Sweezy. How snapchat is using ai and machine learning to thwart drug deals. https://hothardware.com/news/how-snapchat-is-using-ai-and-machine-learning-to-thwart-drug-deals, 2022.

[75] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, et al. Sok: Hate, harassment, and the changing landscape of online abuse. In *IEEE Symposium on Security and Privacy*, pages 247–267, 2021.

[76] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. Towards understanding and detecting cyberbullying in real-world images. In *Network and Distributed System Security Symposium*, 2021.

[77] Ky Vu, Claudia D'Ambrosio, and Leo Liberti. Surrogate-based methods for black-box optimization: Surrogate-based methods for black-box optimization. *International Transactions in Operational Research*, 24, 2016.

[78] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3281–3289, 2021.

[79] Yilin Wang, Jiliang Tang, Jundong Li, Baoxin Li, Yali Wan, Clayton Mellina, Neil O'Hare, and Yi Chang. Understanding and discovering deliberate self-harm content in social media. In *International Conference on World Wide Web*, pages 93–102, 2017.

[80] Lei Xian, Samuel Dakota Vickers, Amanda L Giordano, Jaewoo Lee, In Kee Kim, and Lakshmish Ramaswamy. # selfharm on instagram: Quantitative analysis and classification of non-suicidal self-injury. In *IEEE International Conference on Cognitive Machine Intelligence*, pages 61–70, 2019.

[81] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium*, 2018.

[82] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *Conference on Neural Information Processing Systems*, volume 33, pages 8588–8601, 2020.

[83] Yijun Yang, Ruiyuan Gao, Yu Li, Qiuxia Lai, and Qiang Xu. What you see is not what the network infers: Detecting adversarial examples based on semantic contradiction. In *Network and Distributed System Security Symposium*, 2022.

[84] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014.

[85] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[86] Yue Zhou, Xiaofang Hu, Jiaqi Han, Lidan Wang, and Shukai Duan. High frequency patterns play a key role in the generation of adversarial examples. *Neurocomputing*, 459:131–141, 2021.

## A API Threshold Details

Due to the imbalance of the distribution of the probabilities returned by sexually explicit image detection API, we cannot use 0.5 as the threshold for classifying images into sexually explicit versus non-sexually explicit for Clarifai, Yahoo NSFW, and Amazon Rekognition. We use the GHOST method described by Esposito et al. [32] to find the ideal threshold for the experiment in Table 2. To determine this threshold we randomly drew multiple subsamples from the training data of "porn" class (n = 381) and from the "sexy" class (n = 381) from an NSFW dataset [43]. We make use of the "sexy" class as a safe image category in this experiment because the goal is to select a threshold that properly distinguishes between sexually explicit imagery and safe imagery that contains some similar features to sexually explicit content. After the classification scores are returned by the API, a list of thresholds are screened from 0.5 to 0.95 in increment of 0.05 where the Cohen's kappa is computed with the threshold. The threshold that returns the maximum Cohen's kappa is selected as the threshold for predicting the classification. Using this method, the following thresholds for each moderation API is returned: Clarifai (0.815), Yahoo NSFW (0.881), Amazon Rekognition (0.900). From these results, we chose a threshold of 0.8 for Clarifai, and 0.9 for Yahoo NSFW and Amazon Rekognition. For simplicity, the computed threshold was rounded to the nearest tenth. A quick experiment showed that this rounding had no change in the classification of the images, when the rounded threshold was used instead of the exact threshold.

## B Evading State-of-the-Art Detectors Attack Parameters

Square Attack is launched with parameters $\varepsilon$ = 16/255, n_queries = 10,000, n_restart = 1, loss=cross entropy loss, while AutoAttack is launched with parameter $\varepsilon$ = 8/255. For Square Attack combined with Gaussian blur, we attack an image with Square Attack prior to applying a Gaussian blur with parameters of kernel size = 7, and $\sigma$ = 3.

## C User Study with Online Participants

To evaluate the suitability of region based obfuscation for images on OSN's, we conducted a study of 100 social media users on Amazon MTurk.

### C.1 User Study Methodology

We launched two surveys with mutually exclusive participants with each survey concluding with 50 participants per survey. One survey asked for the participants opinions on fully obfuscated images, and the other survey asked for participants opinions on partially obfuscated images. We used two different images that contained potentially sensitive content and created two versions of each image. The first version of the image had the entire image blurred, and the second version of the image had a regional blur over the unsafe region of the image. We required that our survey participants be located in the United States and be social media users. Each participant was awarded $2 for their completion of the survey, and the average completion time of this survey was approximately 12 minutes and 30 seconds. The experimental protocol was approved by our institution's IRB.

The participants were first instructed to watch a 9-minute video [3] that demonstrates the danger of unsafe images on social media. Next the participants were then told the following: "In this portion of the study, you will be asked to complete a survey. The situations, questions, and answers should be considered thoughtfully and carefully. When answering these questions, think about your experience(s) interacting with photos having sensitive content obfuscated, in general." Next, the participants were asked to view and then rate statements about two images. Both of these images are fully blurred, or both are partially blurred, depending on the version of the survey they received. Then we measured Information Sufficiency, Satisfaction and Perceived Social Presence by asking participants to rate three statements about their thoughts on each image from the choices of: Strongly Agree (7), Agree (6), Somewhat Agree (5), Neither Agree Nor Disagree (4), Somewhat Disagree (3), Disagree (2), Strongly Disagree (1). The three statements that they were asked to rate were (1) "The photo provides sufficient information", (2) "The photo is satisfying", and (3) "There is a sense of human contact when I see the photo", respectively for Information Sufficiency, Satisfaction and Perceived Social Presence. We asked the participants to also rate photo satisfaction, since our framework is also used by personnel who review the safe parts of the image which should not be obfuscated.

| Statements | Fully Obfuscated Image | Partially Obfuscated Image | t | df |
|---|---|---|---|---|
| The photo provides sufficient information | 2.33(0.157) | 4.69(0.156) | 10.657*** | 99 |
| The photo is satisfying | 2.05(0.152) | 3.44(0.161) | 6.264*** | 99 |
| There is a sense of human contact when I see the photo | 2.98(0.189) | 4.25(0.171) | 4.977*** | 99 |

Table 7: Social media content obfuscation user experiment. (*** indicates $p < 0.001$).

### C.2 Results

In Table 7 we show the mean values and the standard error of the mean of the responses on both images from the survey questions on the fully blurred and partially blurred images. We also perform a two-sample t-test and found that the images that had a regional obfuscation were rated higher on all three statements by participants when compared to images that had an obfuscation applied to the entire image. This indicates that the social media user experiences may be improved by adopting regional obfuscation as a method of content control.

[3] https://www.youtube.com/watch?v=dbg4hNHsc_8