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

Authors: Mazal Bethany[1], Andrew Seong[1], Samuel Henrique Silva[1], Nicole Beebe[2], Nishant Vishwamitra[2], and Paul Rad[1]

[1]Secure AI and Autonomy Lab, [2]Department of Information Systems and Cyber Security
University of Texas at San Antonio

USENIX Security 2023

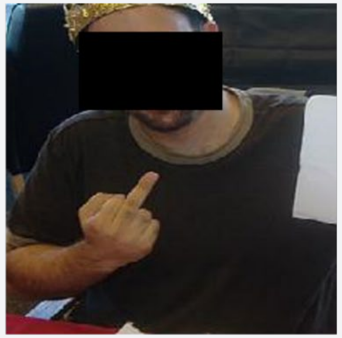
Disclaimer

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

Adversarial Unsafe Images

- **Adversarial Images**: Deceptive digital images that fool AI-based image recognition systems, causing misclassification, while appearing unchanged to human viewers.
- **Unsafe Images**: Potentially harmful or offensive content requiring effective detection and moderation to protect viewers.

NSFW

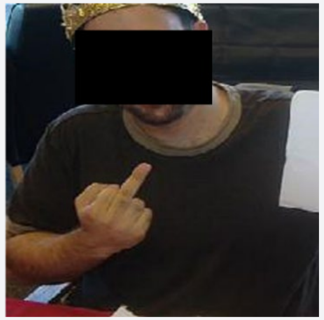


Cyberbullying

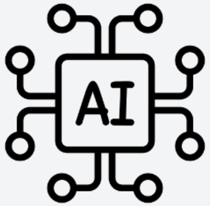
Self-Harm

Detection of Adversarial Unsafe Images

- Small perturbations can fool AI based detectors while preserving visual semantic content

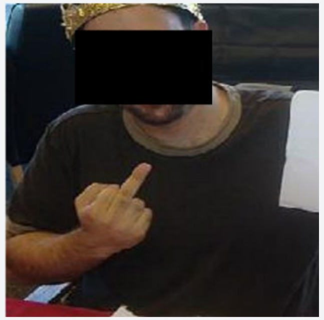


Detected as cyberbullying

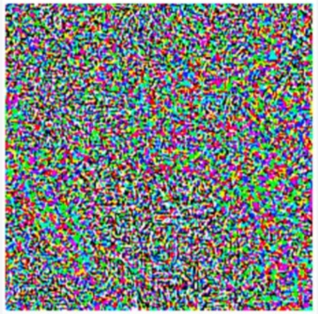


Detection of Adversarial Unsafe Images

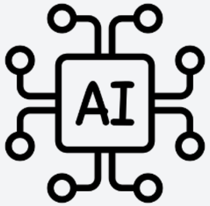
- Small perturbations can fool AI based detectors while preserving visual semantic content



$+ \epsilon *$

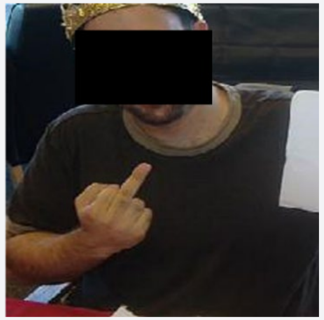


Detected as cyberbullying

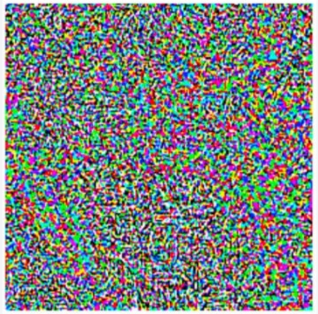


Detection of Adversarial Unsafe Images

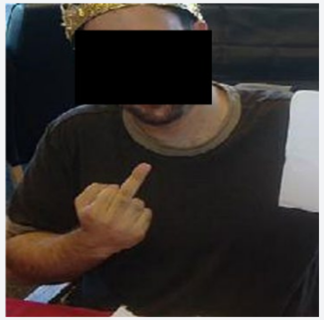
- Small perturbations can fool AI based detectors while preserving visual semantic content



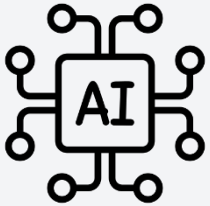
$+ \epsilon *$



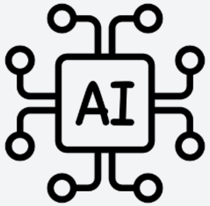
$=$



Detected as cyberbullying



Not detected as cyberbullying



Adversarial Unsafe Images

- Adversarial perturbations compound the issue of unsafe images
- Frequent exposure to unsafe images can cause harm to image reviewers
- Moderator lawsuits for mental damages

How Do Existing Methods Perform?

- With adversarial attacks the detection performance drops almost 40% on average across state-of-the-art API

How Do Existing Methods Perform?

- With adversarial attacks the detection performance drops almost 40% on average across state-of-the-art API

Problem 1: Existing methods are insufficient against adversarial unsafe images

Need a new method that can remove perturbations

Explainability Based Image Obfuscation

- Image obfuscation for protecting reviewers of sensitive images
- Grad-CAM, LIME, Integrated Gradients

Problem 2: Existing explanation methods are unsuitable for image obfuscation

Need new obfuscation methods that are suitable for obfuscation

Motivation Overview

Problem 1: Existing methods are insufficient against adversarial unsafe images → Reconstruction to remove adversarial perturbations

Problem 2: Existing explanation methods are unsuitable for image obfuscation → Counterfactual super region attribution explainability to obfuscate

Datasets

- Sexually Explicit [1]
- Cyberbullying [2]
- Self-Harm
  - Self-harm (self-cutting, self-bruising, eating disorder, depicted or promoted self-harm) (2,100 images)
  - Non-self-harm (neutral social media images) (4,200 images)

[1] Alex Kim. Nsfw data scraper. https://github.com/alex000kim/nsfw_data_scraper, 2021.
[2] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. Towards understanding and detecting cyberbullying in real-world images. In NDSS, 2021.

System Design Intuition

- Reconstruction of Adversarially Perturbed Image with Robust Classifier
  - Image Reconstruction to remove the perturbations as an input transformation defense
  - Robust classifier with adversarial training to detect unsafe content
- Obfuscating Unsafe Content with Counterfactual Explainability
  - Explainability to detect the unsafe parts of the image and obfuscate them

Approach Overview

- uGuard (**u**nsafe image **Guard**)
  - Image reconstruction module: **A**daptive **C**lustering of Robust **S**emantic **R**epresentations (ACSR)
  - Explainability-based image obfuscation module: **C**ounterfactual **S**uper **R**egion **A**ttributions (CSRA)

uGuard Robust model

- **Training a robust image detection model**
  - Adversarial attacks pushes image to tail of training distribution
  - Standard adversarial training:

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

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

uGuard Image Reconstruction

- **Removing Adversarial Perturbation**
  - Reconstruct high-frequency component of image
  - Decompose images into high and low frequency components using the Tikhonov filter
  - Convolutional Dictionary Learning to learn a dictionary from clean (unattacked) images to reconstruct the high frequency component of an image from the low frequency component

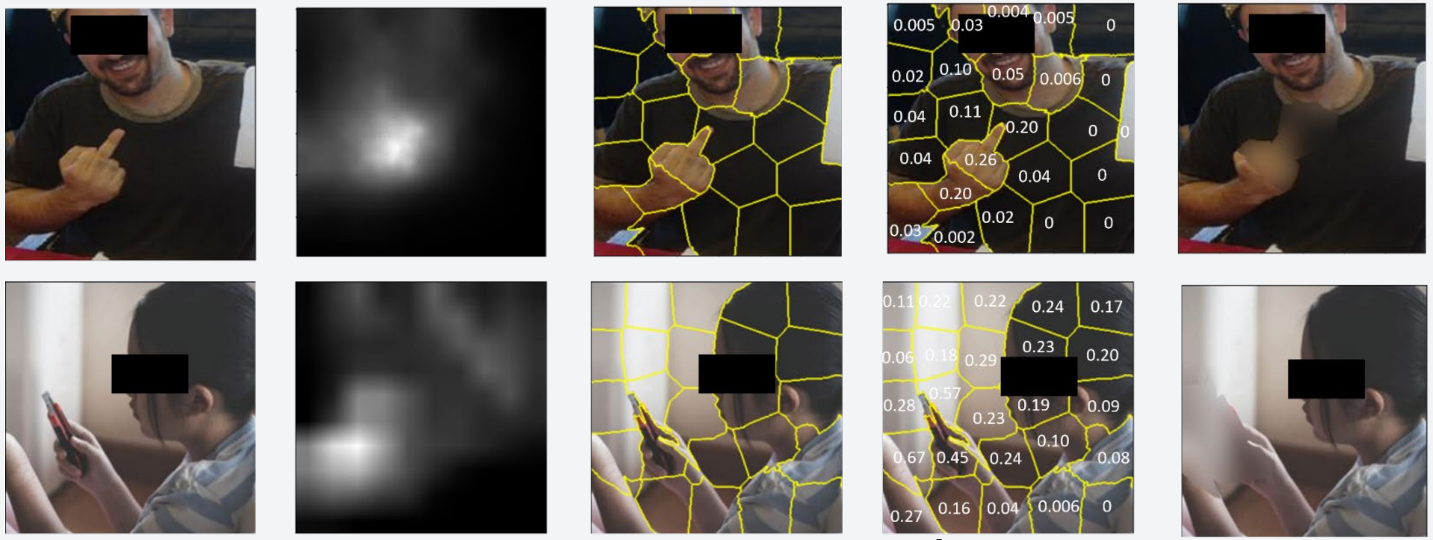
$$x_{rec} = x_{low} + x_{high}^{rec}$$

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

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

uGuard Explainability Based Image Obfuscation

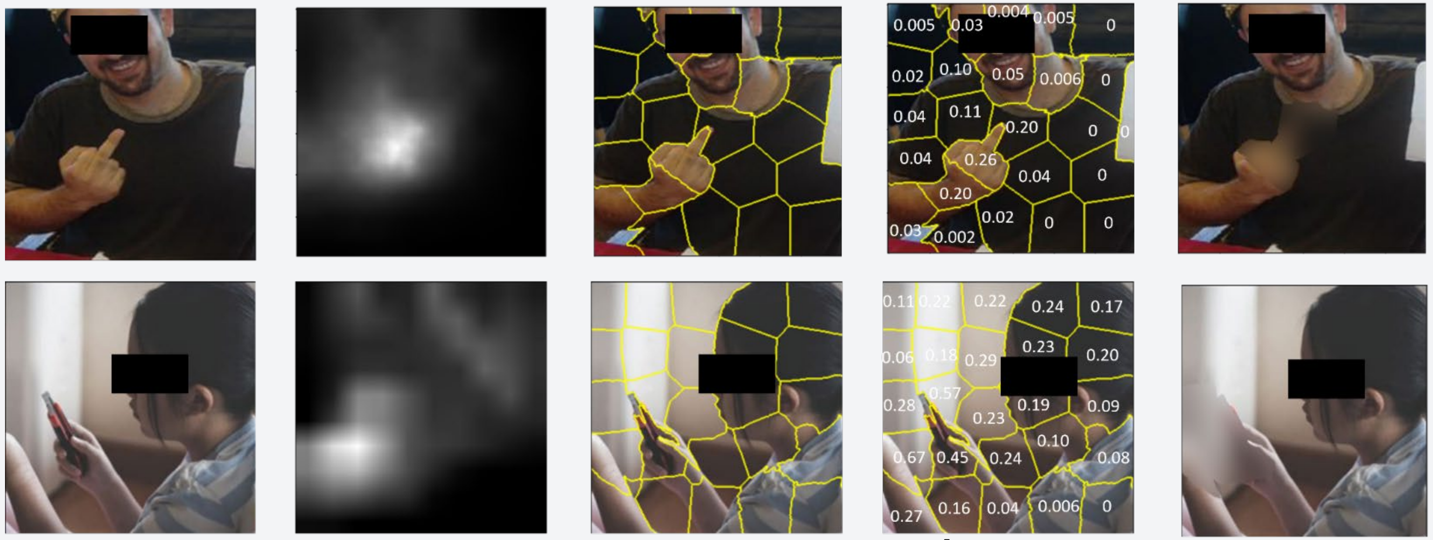
- **Targeted Image Obfuscation**
  - Counterfactual examples
  - 2^K different combination of regions to potentially mask
  - Attribution maps point us to **likely** regions to sample from



Original Image | Attribution Map | Superpixel Segmentation | Average Region Scoring | Ideal Obfuscation Determined by CSRA

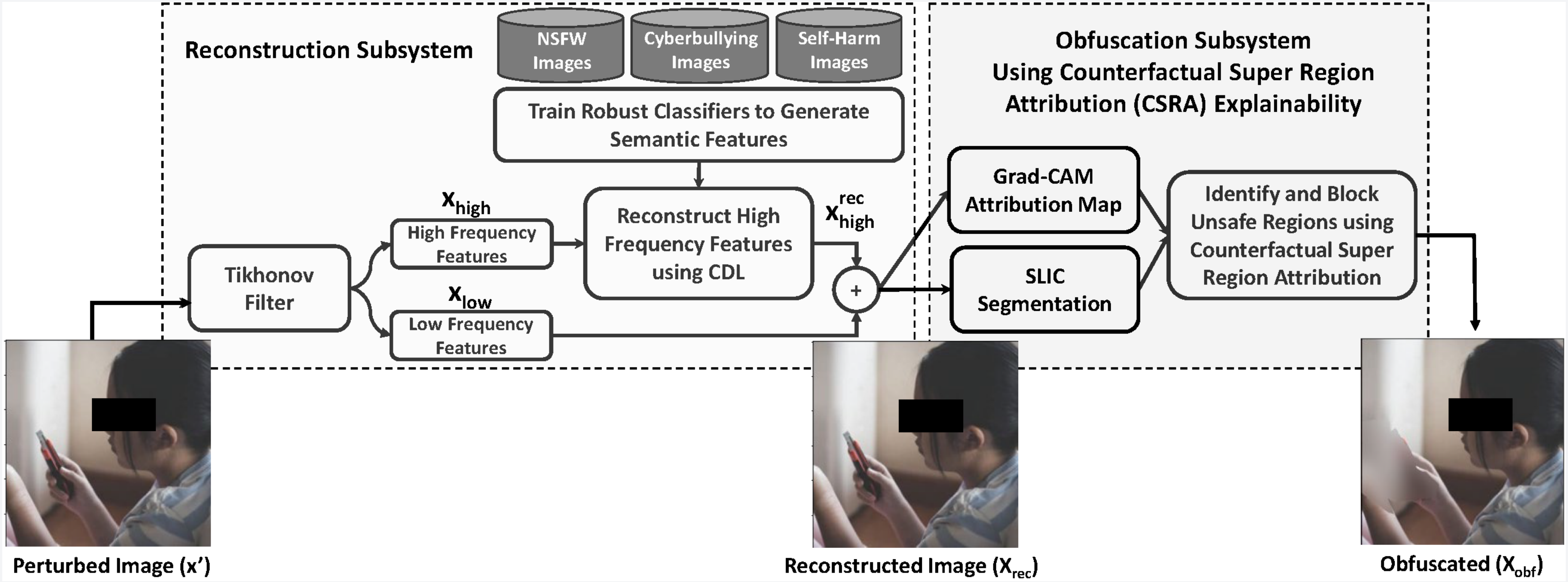
uGuard Explainability Based Image Obfuscation

- Split an image into regions
- Generate attribution map
- Average attribution scores within each segment
- Perform counterfactual analysis of top K scored segments to determine a combination of segments to obfuscate



Original Image | Attribution Map | Superpixel Segmentation | Average Region Scoring | Ideal Obfuscation Determined by CSRA

uGuard System Architecture



Evaluation: Public API vs uGuard

- Public API are unable to perform targeted obfuscation, and perform worse on adversarially perturbed unsafe images

	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

Additional Evaluations

- Adversarial robustness
  - Robustness to seen attacks and some unseen attacks
- Explainability-based obfuscation
  - More images made safer, with less obfuscation overall
  - Preserves more important context than other techniques
- In-the-wild Experiment
  - Human evaluations on sexually-explicit and self-harm images
  - Over 90% of unsafe images made safer

Future Work

- Other unsafe image categories
- Investigating using targeted obfuscation methods in conjunction with Vision Language Models to assist in protecting social media image moderators

Conclusions

- We investigated adversarial unsafe image detection systems and explainability based obfuscation of unsafe images
- State-of-the-art systems that detect unsafe image content are vulnerable to adversarially attacked images
- We presented uGuard to detect and perform targeted obfuscation of adversarial unsafe images across three datasets
- Our evaluations showed that uGuard was able to sufficiently detect and obfuscate adversarially unsafe images

Q&A