Microsoft

edunet foundation

DeepShield: Cloaking Profile Images Against AI Detection

Group Members:

Advait Kulkarni

Soham Shrawankar

Aaryan Tamhane

Aryavardhan Deshmukh

Problem Statement:

Artificial Intelligence (AI) models have achieved remarkable accuracy in facial recognition, enabling them to identify individuals from images with ease. While this technological advancement is impressive, it also poses significant privacy concerns, especially when personal profile images are harvested and analyzed without consent. To address this issue, a promising solution involves adding imperceptible noise—known as adversarial perturbations—to images. These subtle modifications are designed to mislead AI systems while keeping the image visually unchanged to the human eye, thus preserving user privacy without compromising the image's appearance.

Learning Objectives

Understand Adversarial Machine Learning:

- Grasp how AI models can be tricked using adversarial noise or perturbations.
- Learn the principles behind image cloaking and pixel manipulation.

Implement Privacy-Enhancing Techniques:

- Develop methods to protect digital images from unauthorized AI recognition.
- Apply pixel-level defenses without compromising visual quality.

Use Python for Image Processing:

- Utilize libraries such as OpenCV, NumPy, and PIL to manipulate and visualize images.

Develop Tools for Visualizing Adversarial Effects:

- Create interfaces or functions to display cloaked images and invisible layers.

Encourage Ethical AI Usage:

- Reflect on the ethical implications of AI surveillance and user privacy.
- Promote responsible AI usage through defensive mechanisms.

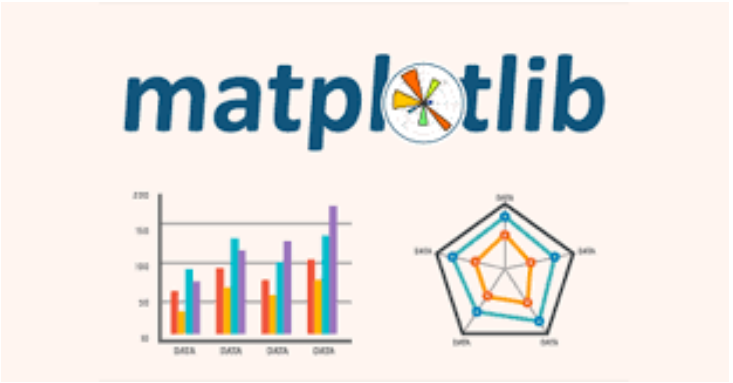
Source : www.freepik.com/

Tools and Technology used

Python libraries: OpenCV, NumPy, PIL,Tkinker , hashlib

Processing includes:

- Noise addition
- Image comparison
- Matrix manipulation



NumPy

Solution:

Effective cloaking techniques offer a powerful way to protect user identity in an increasingly AI-driven digital world. By introducing subtle, carefully crafted modifications to images, these methods can significantly reduce the accuracy of facial recognition systems without noticeably altering the image for human viewers. This ensures that users can continue to use their profile pictures and digital images normally—preserving their visual quality and usability—while simultaneously defending against unauthorized AI-based surveillance or data harvesting. Such an approach not only safeguards personal privacy but also empowers individuals with greater control over their digital presence, striking a balance between security and user experience.

Methodology

Pixel Perturbation

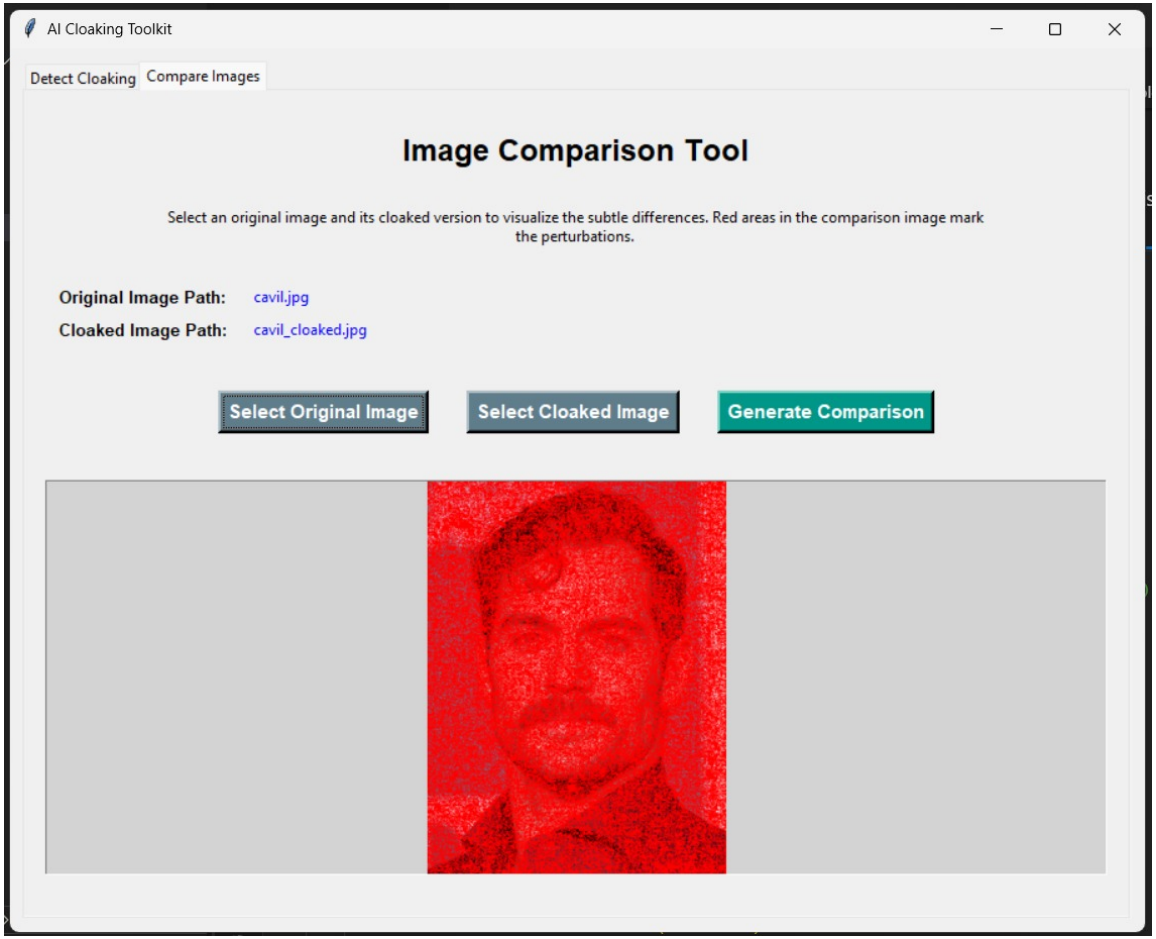
- Load and normalize profile images
- Add tiny, imperceptible noise to confuse AI
- Ensure image remains unchanged to the human eye

Cloaking Layer

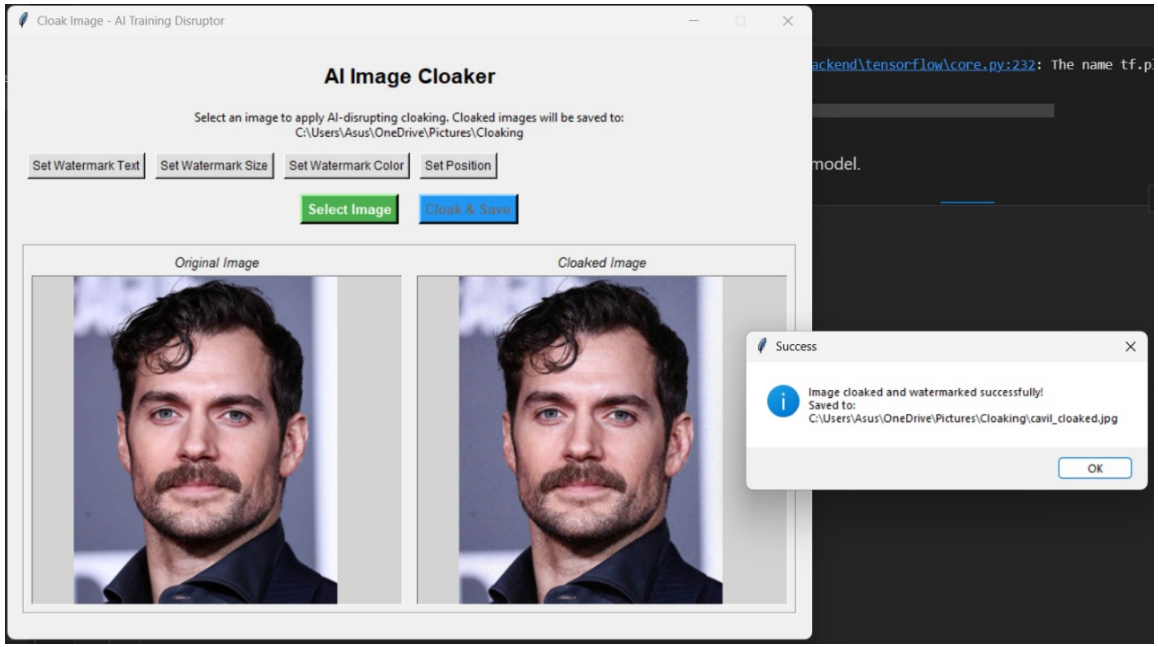
- Generate a noise mask as a protective layer
- Blend it with the original image
- Maintain high visual similarity (using PSNR/SSIM)

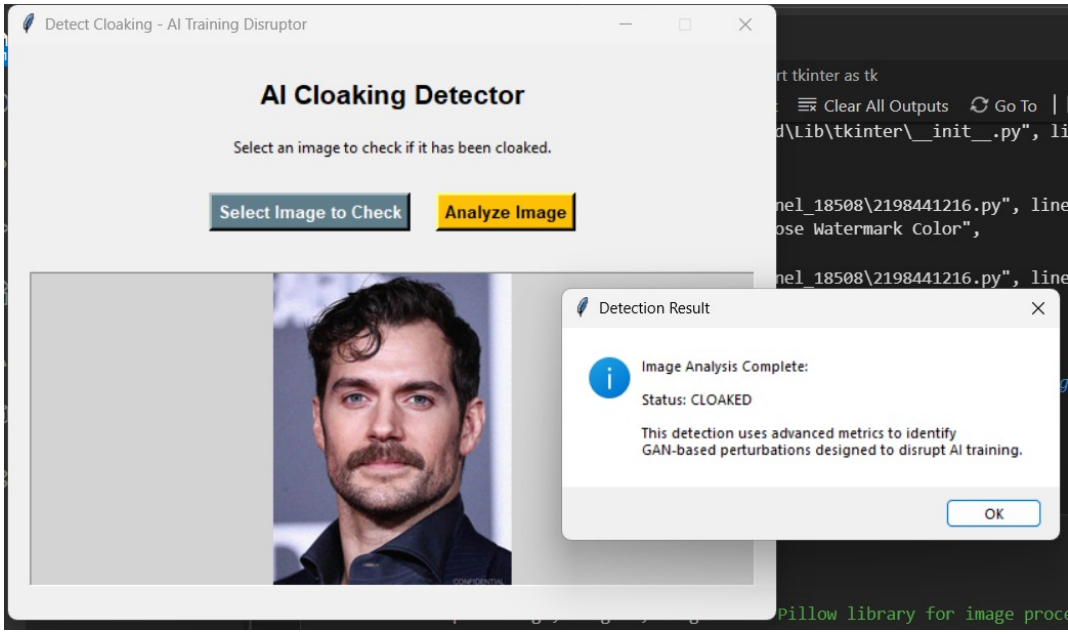
Visualization

- Subtract original from cloaked image
- Amplify and display as a heatmap
- Confirm cloak is invisible to humans but disrupts AI



Screenshot of Output:





Conclusion:

The DeepShield project demonstrates a practical and effective approach to protecting user identity by applying an invisible cloaking layer to profile images. Through subtle pixel perturbations that are unnoticeable to the human eye, the system disrupts the accuracy of AI-based facial recognition models. This ensures that individuals can maintain their online presence without compromising their privacy. The method preserves image quality while providing a strong layer of defense against unauthorized AI surveillance. Visualization techniques further confirm that the cloaking is both discreet and effective, making DeepShield a promising solution for privacy protection in the age of artificial intelligence.