APPLICATION OF ADVERSARIAL MODELS FOR IDENTITY PROTECTION IN IMAGES

Le Minh Tai
University of Information Technology
HCMC, Vietnam

What ?

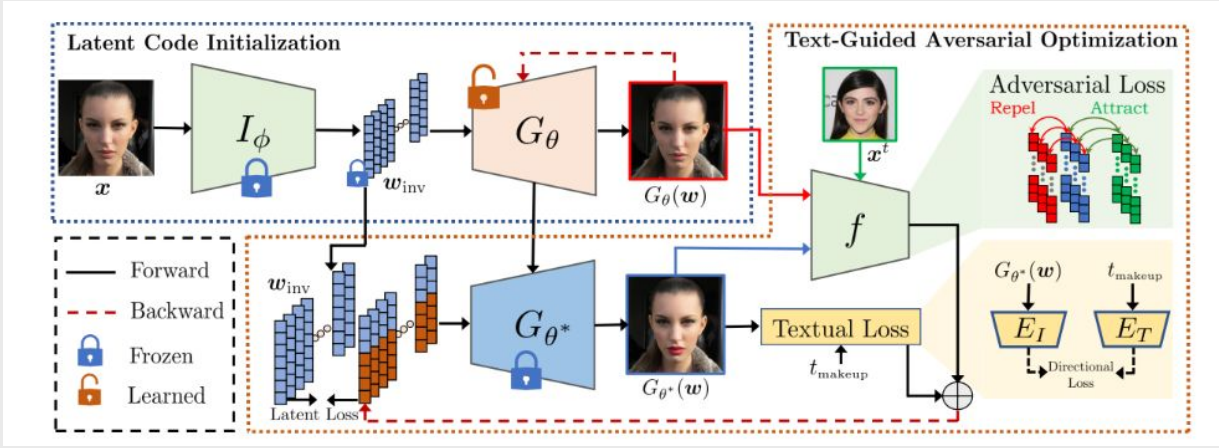
We propose CLIP2Protect, a novel method to protect facial privacy by generating realistic adversarial makeup faces guided by text prompts.
Our approach modifies only the makeup appearance of a face image without altering identity-defining features, to fool black-box facial recognition systems, while still looking natural to humans.

Why ?

- Deep face recognition models are widely used for surveillance and can track users without consent.
- Existing privacy methods often create unnatural artifacts or require reference makeup images, limiting practicality.
- CLIP2Protect offers:
  💬User-controlled privacy via simple text (e.g., "pink eyeshadow")
  🫂Realistic images that match human expectations
  💪Strong protection against black-box face recognition

Overview

Latent Code Initialization → Text-Guided Adversarial Optimization → Final Output



Description

0. Core Idea

CLIP2Protect generates makeup-style face images that:

- Look natural to humans
- Are unrecognizable to face recognition systems (FR)
- Follow user-defined makeup text prompts

Figure 1. Protected images (bottom) are generated from original faces (top) using user-defined makeup prompts.

1. Latent Code Initialization

- The original face image is encoded into a latent code using a pretrained encoder (e4e).
- This code is used to reconstruct the face with a fine-tuned StyleGAN generator to preserve identity.

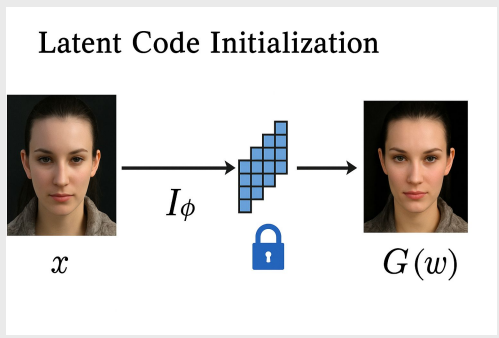


Figure 2. The original face is encoded into a latent code and reconstructed using a fine-tuned generator to preserve identity.

2. Text-Guided Adversarial Optimization

- A makeup text prompt (e.g., “pink eyeshadows”) guides face editing using CLIP.
- Only deep layers (8–18) in latent space are optimized to apply makeup without changing identity.
- The latent code is adjusted to:
  - Match the makeup prompt (Textual Loss)
  - Fool face recognition (Adversarial Loss)
  - Preserve identity in key layers (Latent Loss)

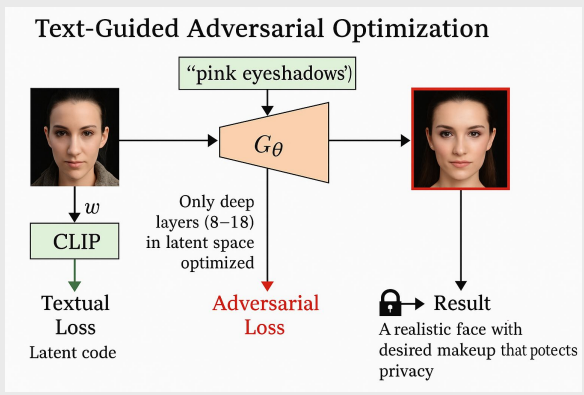


Figure 3. The latent code is optimized using a makeup prompt (e.g., "pink eyeshadows") and CLIP losses to generate a realistic face that fools recognition systems while preserving identity.

3. Final Output

- Realistic face with desired makeup
- Breaks FR matching while retaining human-recognizable identity
- Fully driven by text prompts. no reference image needed

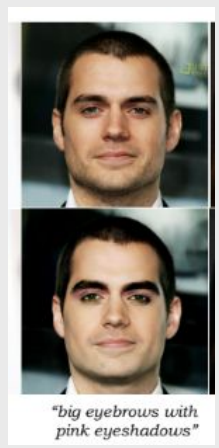


Figure 4. Before and After

NII
Lê Minh Tài – Trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM
TEL : 0373435788 Email : tailm.19@grad.uit.edu.vn