

BẢO VỆ HÌNH ẢNH KHỎI MÔ HÌNH MÁY HỌC BẰNG MIẾNG VÁ ĐỐI KHÁNG NGỤY TRANG TỰ NHIÊN.

*PROTECTING IMAGES FROM MACHINE LEARNING MODELS USING NATURALLY
CAMOUFLAGED ADVERSARIAL PATCHES.*

Huỳnh Chí Nhân- 22520996

Nguyễn Hồ Nam – 22520915

Tóm tắt



Huỳnh Chí Nhân
22520996

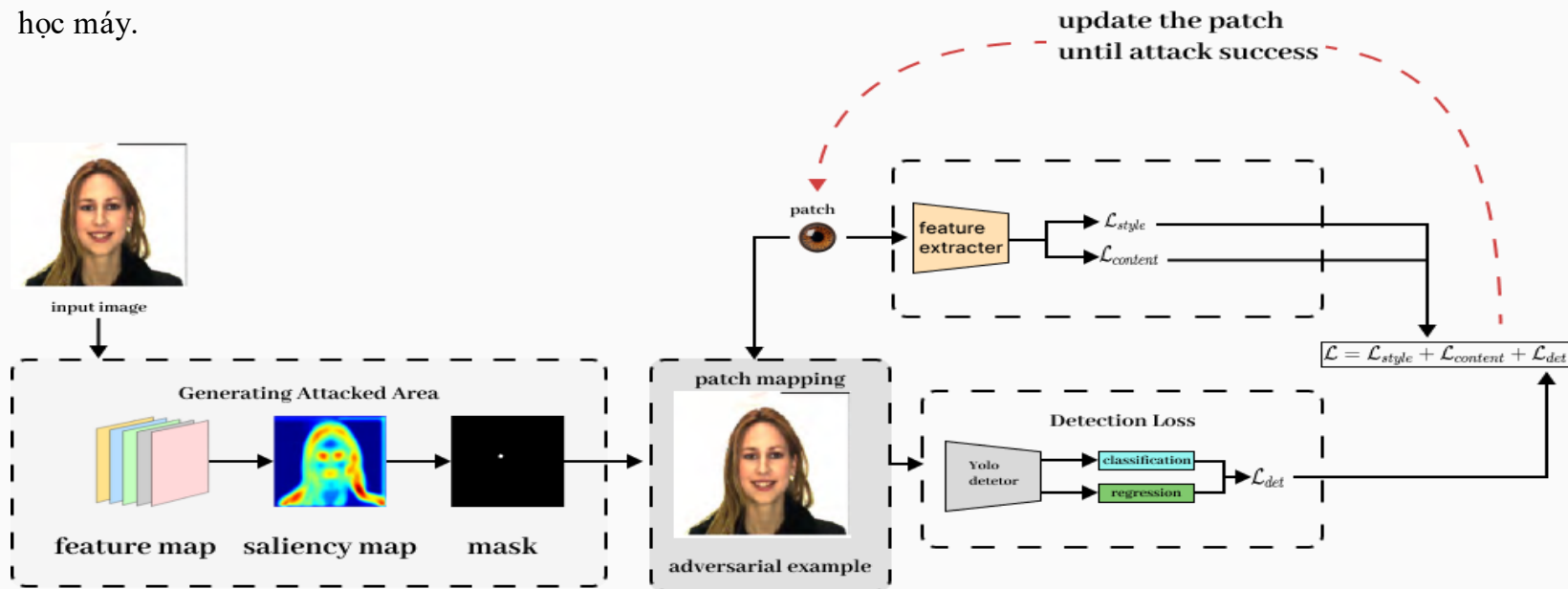


Nguyễn Hồ Nam
22520915

- Lớp: CS519.P11
- Link Github của nhóm: <https://github.com/nhanhuynh123/CS519.P11>
- Link YouTube video:

Giới thiệu

- **Deepfake** là một mô hình trí tuệ nhân tạo mạnh mẽ, có khả năng tái tạo hình ảnh, video, và âm thanh tương tự như của một cá nhân bất kỳ. **Deepfake** mang tính giải trí cao nhưng cũng tiềm ẩn nguy cơ bị lạm dụng.
- Tận dụng tính nhạy cảm với tấn công đối kháng của mô hình học sâu dùng **DNNs**, chúng tôi đề xuất sử dụng **Camouflaged Adversarial Patches** - một kiểu tấn công đối kháng để bảo vệ dữ liệu ảnh chống lại các mô hình học máy.



Mục tiêu

1. Thử nghiệm và đánh giá hiệu quả của **Adversarial Patches** theo định hướng **Digital Attack** trên mô hình **YOLO**.
2. Cải thiện tính tự nhiên của **Adversarial Patches** bằng kỹ thuật **Style Transfer**, đồng thời duy trì **ASR** ổn định.
3. Thiết kế và đóng gói mô hình **Camouflaged Adversarial Patches** dưới dạng module, hướng đến ứng dụng trong các hệ thống bảo mật dữ liệu ảnh cá nhân và tổ chức.

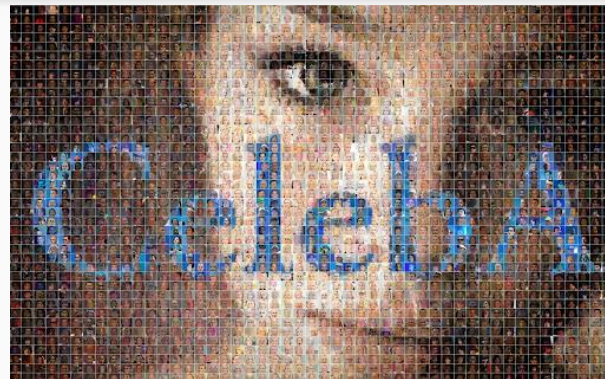
Nội dung và Phương pháp

Nội dung 1: Tìm hiểu lý thuyết về Adversarial Attack và cài đặt môi trường.

Hiểu rõ hoạt động của **Digital Attack**.

Cài đặt môi trường:

- Sử dụng bộ dữ liệu **CelebA**, gồm hình ảnh khuôn mặt người.
- Chọn mô hình phân loại **YOLOv8** làm mục tiêu tấn công.



Hình 1: Bộ dữ liệu CelebA.



Hình 2: Mô hình phát hiện YOLOv8.

Nội dung và Phương pháp

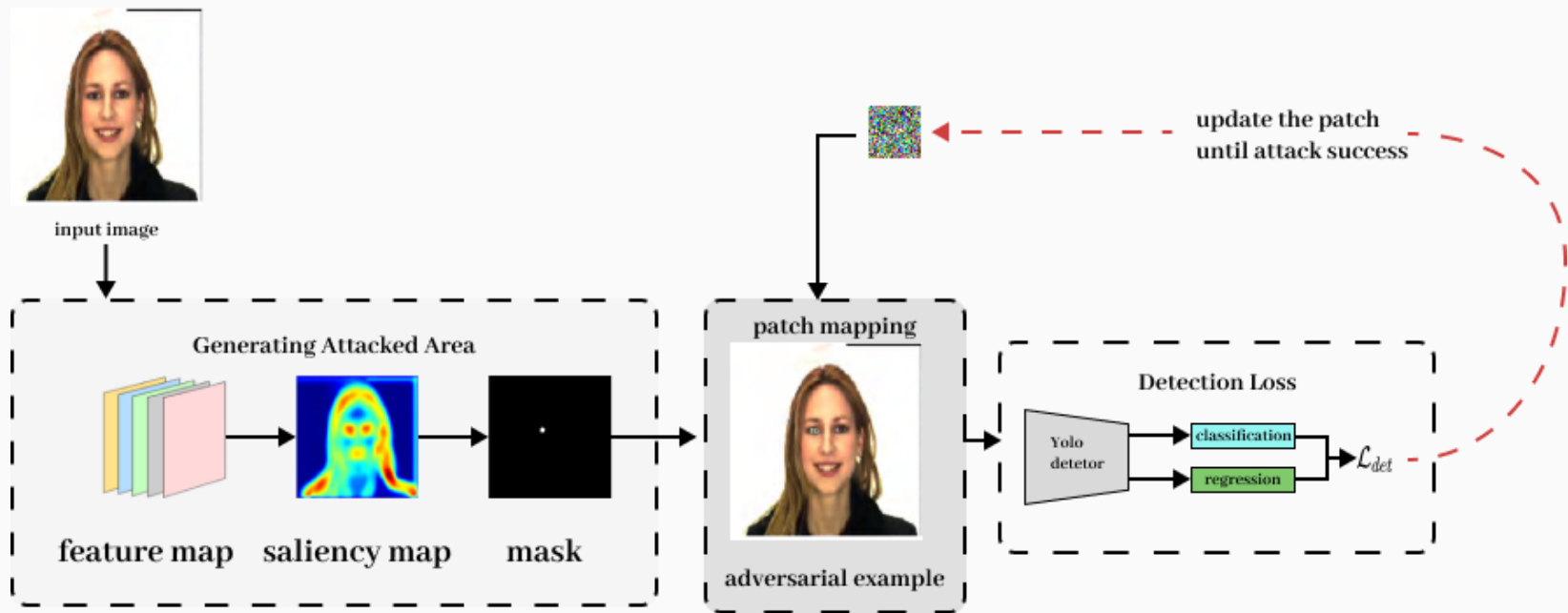
Nội dung 2: Thử nghiệm đánh giá hiệu năng của Adversarial Patch.

Cài đặt Adversarial Patch:

- Sử dụng **Grad-CAM++** [5] để tạo **Saliency Map**, xây dựng **Mask** xác định vị trí, kích thước của miếng vá (**Patch**) tại khu vực quan trọng đối với dự đoán của mô hình.
- Phát sinh **Adversarial Patch** theo vị trí, kích thước được chỉ định bởi **Mask**.
- Patch được tối ưu qua từng tập huấn luyện, bằng cách tối ưu hàm **Loss = Loss Detection**, từ đó miếng vá làm lệch dự đoán của mô hình về bounding box, class của vật thể trong ảnh [3].

Kiểm tra hiệu quả tấn công của **Adversarial Patches** trên bộ dữ liệu **CelebA** và mô hình **YOLO**.

Nội dung và Phương pháp



Hình 3: Sơ đồ hoạt động của Aversarial Patches.

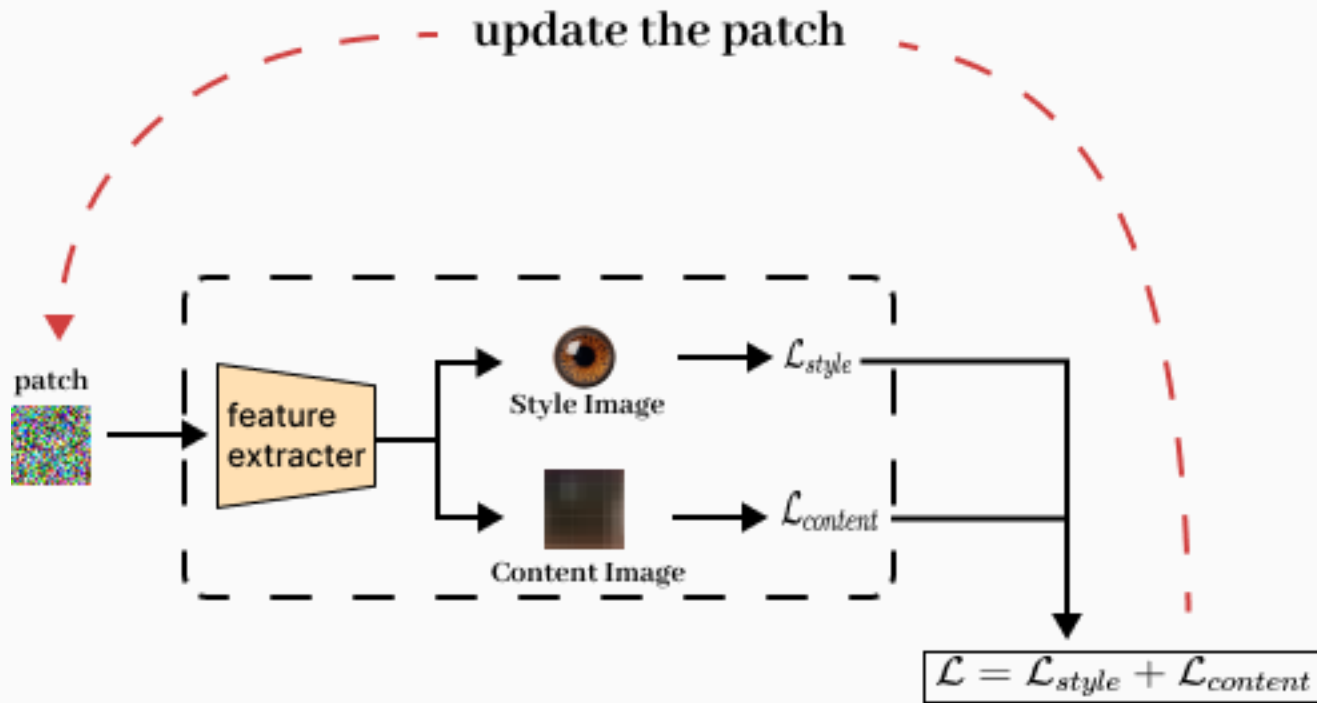
Nội dung và Phương pháp

Nội dung 3: Tiến hành nguy trang Patch sử dụng Style Transfer.

Phương pháp:

- Lựa chọn **Style Image** và **Content Image**.
- Xây dựng hàm **Loss = Content Loss + Style Loss**:
 - ❖ **Content Loss**: Đo độ khác biệt về nội dung giữa **Content Image** và **Patch**.
 - ❖ **Style Loss**: Đo độ khác biệt về phong cách giữa **Style Image** và **Patch**.
- Tối ưu hàm mất mát để làm tăng độ tương đồng giữa **Patch** với **Style Image** và **Content Image**.

Nội dung và Phương pháp



Hình 4: Sơ đồ hoạt động của kỹ thuật Style Transfer.

Nội dung và Phương pháp

Nội dung 4: Thử nghiệm, đánh giá hiệu năng của Camouflaged Adversarial Patches.

Phương pháp:

Thử nghiệm CAP trên mô hình phân loại **YOLO**.

Đánh giá thông qua các tiêu chí: **ASR (Attack Success Rate)**, **SSIM (Structure Similarity Index Measure)**, và **Patch Size**.

Tiêu chí đánh giá:

- **ASR (Attack Success Rate):** Xác suất thành công của adversarial patch trong việc làm mô hình dự đoán sai.
- **SSIM (Structure Similarity Index Measure):** Đo độ tương đồng giữa ảnh gốc và Adversarial Examples (ảnh đã thêm Adversarial Patch).
- **Patch Size:** kích thước của miếng vá (Patch).

Kết quả dự kiến

- 1. Huấn luyện thành công mô hình Adversarial Patch, ASR trung bình đạt trên 80% khi tấn công mô hình YOLO.
- 2. Phát sinh thành công mô hình Camouflaged Adversarial Patches. Thử nghiệm tấn công mô hình YOLO và đạt SSIM trên 0.8, ASR trên 70%
- 3. Mô hình ổn định và được đóng gói thành module.

Tài liệu tham khảo

- [1]. Gan Pei, Jiangning Zhang, Menghan Hu, Guangtao Zhai, Chengjie Wang, Zhenyu Zhang, Jian Yang, Chunhua Shen and Dacheng Tao, "Deepfake Generation and Detection: A Benchmark and Survey," CoRR, vol. abs/2403.17881, 2024.
- [2]. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras and Adrian Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," CoRR, vol. abs/1706.06083, 2017.
- [3]. Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li and Yiran Chen, "DPatch: An Adversarial Patch Attack on Object Detectors," in AAAI, 2019.
- [4]. Ranjie Duan, Xingjun Ma, Yisen Wang, James Bailey, A. Kai Qin and Yun Yang, "Adversarial Camouflage: Hiding Physical-World Attacks With Natural Styles," in IEEE/CVF, 2020, pp. 997-1005.
- [5]. Chattopadhyay, Aditya and Sarkar, Anirban and Howlader, Prantik and Balasubramanian and Vineeth, "Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks," in WACV, 2018, pp. 839-847.