TRƯỜNG ĐẠI HỌC BÁCH KHOA

KHOA CÔNG NGHỆ THÔNG TIN

PBL6: DỰ ÁN CHUYÊN NGÀNH

KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO

HỆ THỐNG NHẬN DẠNG ẢNH MẶT NGƯỜI TẠO RA BỞI TRÍ TUỆ NHÂN TẠO

GIẢNG VIÊN HƯỚNG DẪN: TS. Ninh Khánh Duy

HỌ VÀ TÊN SINH VIÊN	LỚP SINH HOẠT	LỚP HỌC PHẦN
Cao Kiều Văn Mạnh	20TCLC_KHDL	20.15B
Võ Hoàng Bảo		
Nguyễn Tuấn Hưng		

ĐÀ NẴNG, 12/2023

BẢNG PHÂN CÔNG NHIỆM VỤ

Sinh viên thực hiện	Các nhiệm vụ	Đánh giá
Cao Kiều Văn Mạnh	Phân công công việc, đảm bảo tiến độ thực hiện đồ án	Hoàn thành
	Tìm hiểu, triển khai mô hình học sâu	
	Thử nghiệm các phương pháp cải thiện độ chính xác mô hình	
	Viết báo cáo, chuẩn bị nội dung thuyết trình	
Võ Hoàng Bảo	Thu thập và xây dựng bộ dữ liệu	Hoàn thành
	Thực hiện quy trình làm sạch và tiền xử lý dữ liệu	
	Tích hợp các phương pháp giải thích kết quả dự đoán của mô hình vào hệ thống	
	Viết báo cáo, chuẩn bị nội dung thuyết trình	
Nguyễn Tuấn Hưng	Xây dựng server	Hoàn thành
	Xây dựng website dịch vụ	
	Tích hợp mô hình vào hệ thống	
	Viết báo cáo, chuẩn bị nội dung thuyết trình	

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

TÓM TẮT ĐỒ ÁN

Với sự phát triển mạnh mẽ của trí tuệ nhân tạo (AI), theo sau đó là các mô hình trí tuệ nhân tạo tạo sinh ảnh, việc phân biệt hình ảnh do AI tạo ra và hình ảnh thật ngày càng khó khăn. Sự phát triển của công nghệ trên đã đặt ra nhiều mối nguy hại về độ tin cậy và an toàn trong nhiều lĩnh vực như bảo mật thông tin, quyền riêng tư, quyền sở hữu trí tuệ,... Điều này còn đặc biệt quan trọng đối với các dữ liệu liên quan đến định danh như hình ảnh khuôn mặt con người. Hiện nay, tuy đã có nhiều nghiên cứu nhằm phân biệt hình ảnh thực tế và hình ảnh do trí tuệ nhân tạo sinh ra, tuy nhiên các mô hình vẫn chưa đạt được kết quả quá cao với dữ liệu từ các mô hình sinh ảnh mới, chưa quá tập trung vào mảng dữ liệu khuôn mặt con người, cũng như vẫn chưa được áp dụng vào trong một hệ thống thực tế.

Do đó, trong nghiên cứu này, nhóm tiến hành xây dựng một bộ dữ liệu ảnh mặt người do AI tạo ra từ các mô hình sinh ảnh mới nhất. Bên cạnh đó, nhóm cũng đề xuất một số mô hình học sâu dựa trên các kiến trúc Convolutional Neural Network và các kỹ thuật Transfer Learning - Ensemble Learning nhằm phát hiện ảnh mặt người thực tế và ảnh mặt người do trí tuệ nhân tạo tạo ra. Ngoài ra, nhóm cũng tiến hành xây dựng hệ thống website để tích hợp kết quả nghiên cứu, giúp người dùng có thể dễ dàng tương tác và sử dụng.

Qua quá trình thử nghiệm và đánh giá trên tập dữ liệu nhóm xây dựng, kết quả cho thấy sự hiệu quả của các mô hình với độ chính xác cao nhất lên đến 97.91% cùng với hệ thống có tốc độ xử lý nhanh, trung bình xử lý 7 - 8s/ảnh cho quá trình phân loại và khoảng 70 - 74s/ảnh cho quá trình giải thích kết quả phân loại. Tuy nhiên, kết quả vẫn còn nhiều thiếu sót và hạn chế. Nhóm chúng em sẽ tiếp tục phát triển và hoàn thiện hơn trong tương lai.

1. TỔNG QUAN ĐỀ TÀI

1.1. Vấn đề và tính cấp thiết của dự án

Trong thời đại của Trí tuệ nhân tạo và Generative AI, việc phân biệt hình ảnh do AI tạo ra và hình ảnh thật ngày càng khó khăn. Các mô hình hiện nay đã có thể sinh ra các hình ảnh vô cùng chân thực và gần như không thể xác định bằng mắt thường mà cần phải dùng đến các mô hình học sâu để phân biệt.

Sự phát triển của công nghệ trên đã đặt ra nhiều vấn đề về độ tin cậy và an toàn trong nhiều lĩnh vực như xác minh danh tính trực tuyến và bảo vệ quyền riêng tư, quyền sở hữu trí tuệ v.v.. Vấn đề này càng đặt biệt quan trọng trong các lĩnh vực xác thực thông tin như xác thực khuôn mặt người. Khuôn mặt con người là một phần quan trọng trong việc định danh cá nhân. Bên cạnh đó, thông tin khuôn mặt còn đóng vai trò quan trọng trong nhiều lĩnh vực khác như thanh toán di động, giao dịch trực tuyến, xác thực mật khẩu v.v.. Nhóm đã tiến hành tìm hiểu và nhận thấy hiện tại có khá ít các mô hình, sản phẩm cung cấp các dịch vụ phân loại hình ảnh do AI tạo ra. Bên cạnh đó, độ chính xác của các sản phẩm hiện có vẫn chưa quá tốt hoặc là các dịch vụ tính phí cao. Đặc biệt, các mô hình hiện có thường không tập trung vào một lĩnh vực cụ thể như khuôn mặt người do đó dẫn đến kết quả chưa quá cao.

Từ thực tế trên, nhóm chúng em tiến hành nghiên cứu và phát triển đề tài "*Hệ thống nhận dạng ảnh mặt người tạo ra bởi trí tuệ nhân tạo (A system for detecting AI generated human faces images)*" nhằm giải quyết các vấn đề trên.

1.2. Mục tiêu đề tài

Các mục tiêu chính của dự án lần này bao gồm: (1) xây dựng thành công bộ dữ liệu mặt người do AI tạo ra đa dạng và có chất lượng cao; (2) thử nghiệm và xây dựng các mô hình, kỹ thuật học sâu để phân loại được hình ảnh mặt người do AI tạo ra và hình ảnh trong thực tế và (3) tiến hành tích hợp mô hình trên vào hệ thống website và cung cấp dịch vụ cho người dùng một cách chính xác, đơn giản và hiệu quả.

2. CƠ SỞ LÝ THUYẾT

2.1. Ý tưởng

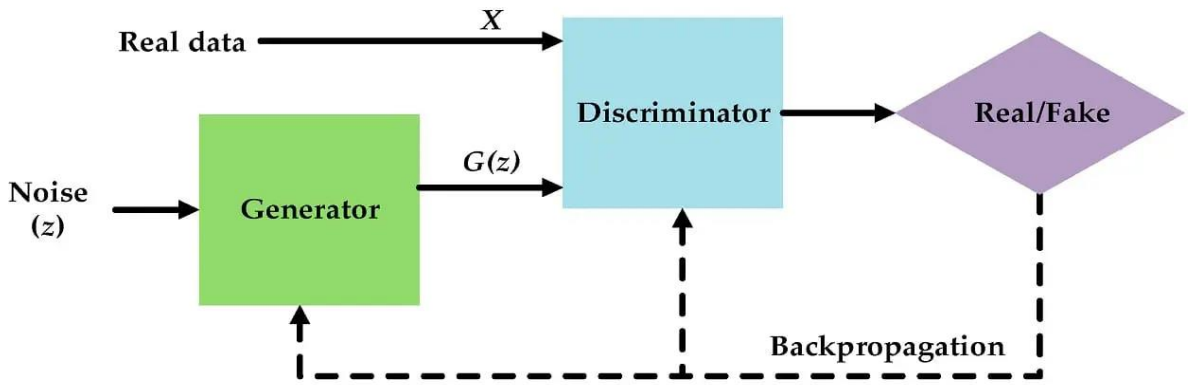
Các bức ảnh do AI tạo ra dù có chân thật đến đâu cũng có những điểm không hài hòa hoặc bất cân xứng. Nắm bắt nhược điểm đó, bằng cách xây dựng một bộ dữ liệu đủ đa dạng và sử dụng các mô hình/kỹ thuật học sâu, chúng ta có thể xây dựng được một mô hình phát hiện ảnh mặt người thực tế và ảnh do trí tuệ nhân tạo tạo ra. Và từ đó, tích hợp mô hình vào hệ thống dịch vụ để cung cấp cho người dùng.

2.2. Các mô hình sinh ảnh hiện nay

Hầu hết các mô hình sinh ảnh hiện nay đều được xây dựng dựa trên ***Kiến trúc GANs***. Nhìn chung, kiến trúc GANs bao gồm 2 mạng chính:

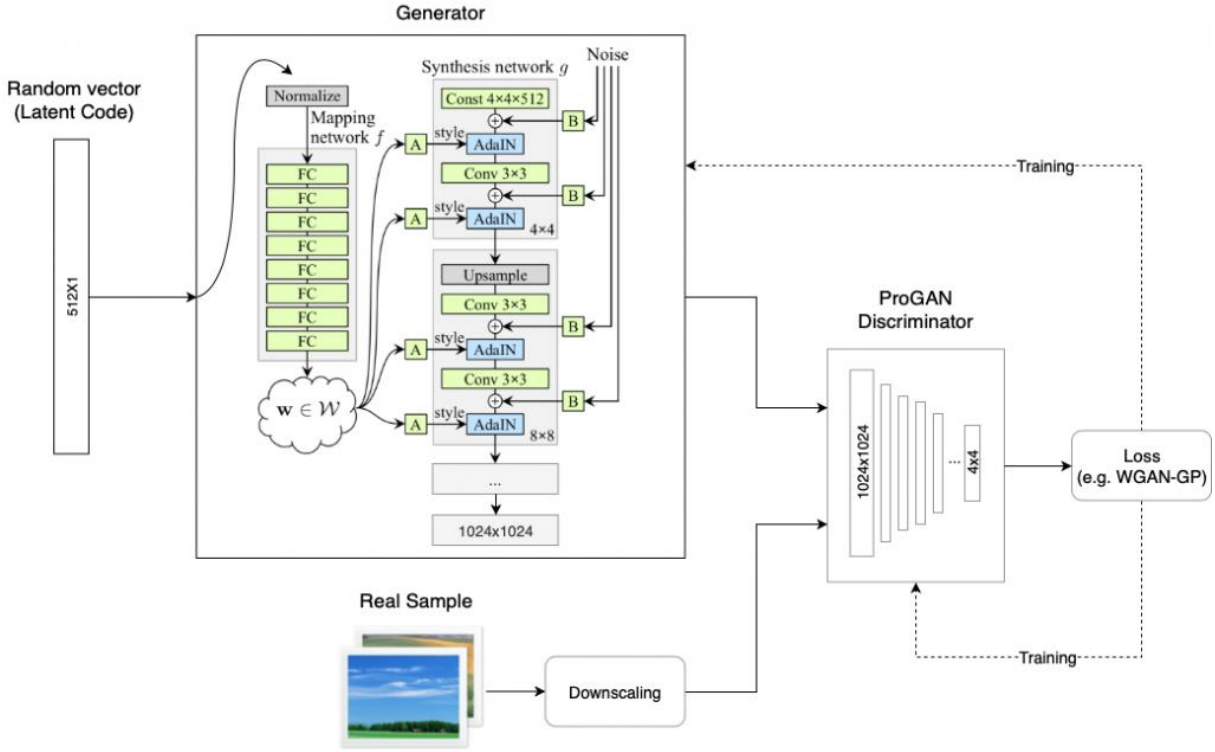
- ***The generator network*** với mục tiêu là sản xuất ra các ảnh khó phân biệt so với dữ liệu thực tế.
- ***The discriminator network*** với mục tiêu cố gắng phân biệt giữa dữ liệu thật được đưa vào huấn luyện và dữ liệu được sinh ra bởi generator.

Một số biến thể của GANs có thể kể đến là: DCGANs, CGANs, CycleGANs, StyleGANs,...



Hình 1. Sơ đồ hoạt động kiến trúc GANs

Trong đó, ***StyleGANs*** [13] là mô hình sinh ảnh có chất lượng chi tiết tốt nhất và là mô hình phổ biến nhất được sử dụng để sinh ảnh mặt người giả. Bên cạnh đó, StyleGAN cũng thường được dùng để cải thiện ảnh mặt người dựa trên hình ảnh ban đầu sinh ra từ các mô hình khác như Diffusion.



Hình 2. Kiến trúc StyleGANs

2.3. Các phương pháp nhận diện ảnh do trí tuệ nhân tạo sinh ra

Các nghiên cứu dùng để nhận diện ảnh do Trí tuệ nhân tạo sinh ra đã và đang được phát triển bằng nhiều phương pháp khác nhau. Hiện nay, các phương pháp này được chia thành ba hướng nghiên cứu chính: dựa trên các kỹ thuật thị giác máy tính (Computer Vision), dựa trên các mạng học sâu CNN và dựa trên các kiến trúc học sâu riêng biệt của một số phương pháp mới gần đây.

2.3.1. Các phương pháp sử dụng Computer Vision

Với phương pháp dựa trên các kỹ thuật Computer Vision sẽ chủ yếu sử dụng các đặc trưng nông của ảnh như màu sắc, tần số và sử dụng các kỹ thuật Computer Vision như Histogram Color,... để tiến hành phân loại.

Trong nghiên cứu [8], tác giả đã tiến hành phân tích thống kê thành phần màu của ảnh ảo và phân biệt với ảnh thật. Hay như trong nghiên cứu [9], tác giả nhận diện dựa trên màu sắc và độ bão hòa. Họ cho rằng dựa trên màu sắc, ảnh do AI sinh ra sẽ có độ tương quan giữa các pixel với nhau trong không gian sắc độ khác với ảnh thật.

Điểm mạnh của các phương pháp trên là mô hình nhỏ, nhẹ, tốc độ xử lý nhanh và đạt được kết quả phân loại tốt với các mô hình sinh ảnh đời đầu. Tuy nhiên, với các mô hình, kiến trúc sinh ảnh mới gần đây thì các phương pháp chỉ dựa trên Computer Vision gần như không thể xử lý tốt.

2.3.2. Các phương pháp sử dụng kiến trúc CNN

Các mạng học sâu CNN đã được áp dụng thành công trong nhiều lĩnh vực, bao gồm cả bài toán phân loại ảnh thật và ảnh do AI sinh ra. Các phương pháp này thường sử dụng các mô hình CNN sâu để học các đặc trưng phức tạp của ảnh, từ đó có thể phân biệt được ảnh giả.

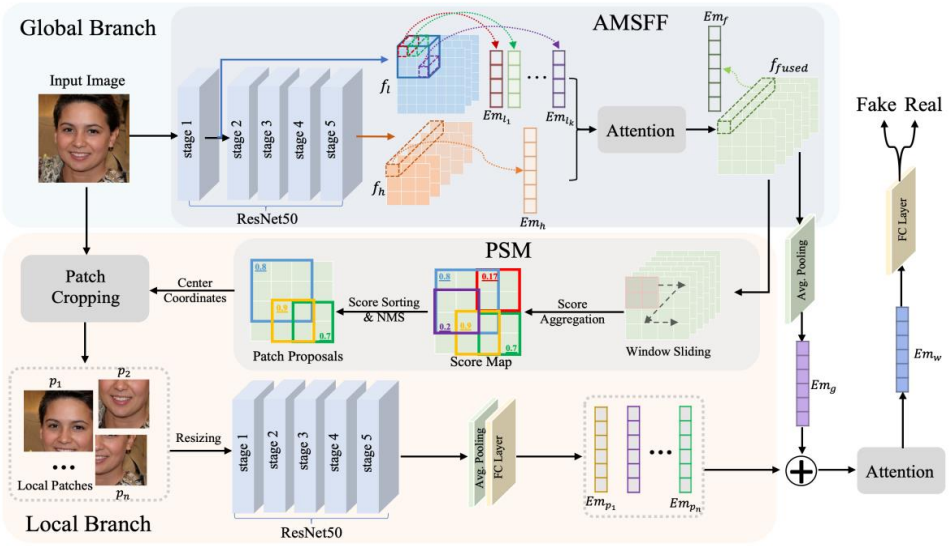
Trong nghiên cứu [10], các tác giả đã sử dụng mô hình học máy phân loại dựa trên ResNet18 để phát hiện hình ảnh được tạo bằng Stable Diffusion với prompt (chuyển văn bản thành hình ảnh). Cụ thể, họ đã thay đổi một số thành phần của ResNet18 để cải thiện khả năng phát hiện hình ảnh giả.

Các phương pháp sử dụng kiến trúc CNN thường có độ chính xác cao hơn các phương pháp chỉ sử dụng Computer Vision. Tuy nhiên, các phương pháp này thường yêu cầu lượng dữ liệu huấn luyện lớn và được cập nhật thường xuyên để đảm bảo độ chính xác cao trên các mô hình sinh ảnh mới.

2.3.3. Các phương pháp gần đây

Gần đây, một số phương pháp mới đã được đề xuất để giải quyết bài toán phân loại ảnh thật và ảnh do AI sinh ra. Các phương pháp này thường sử dụng các kỹ thuật mới như học sâu, học máy tự động, v.v... để cải thiện độ chính xác và hiệu quả của hệ thống.

Cụ thể, trong nghiên cứu [1], tác giả đã kết hợp đặc trưng lớp convolutional nông và sâu của một bức ảnh (nhánh global) cùng với các đặc trưng của các patch được chia nhỏ từ bức ảnh (nhánh local). Ở nhánh global, tác giả sử dụng kĩ thuật "Attention-based Multi-Scale Feature Fusion (AMSFF)" để trích xuất các đặc trưng tổng thể như: màu sắc chung của ảnh, tổng cộng độ sáng, tỷ lệ khung hình,... Tiếp theo, tác giả sử dụng kỹ thuật Patch Selection Module để chọn lọc các khung ảnh quan trọng để trích xuất đặc trưng cục bộ như sự mất cân đối hoặc sai lệch của các chi tiết. Đặc trưng của cả hai nhánh sau cùng được kết hợp lại bởi Attentional Feature Fusion Module và được đưa vào hệ thống phân loại nhị phân để phân biệt ảnh giả và ảnh thật.



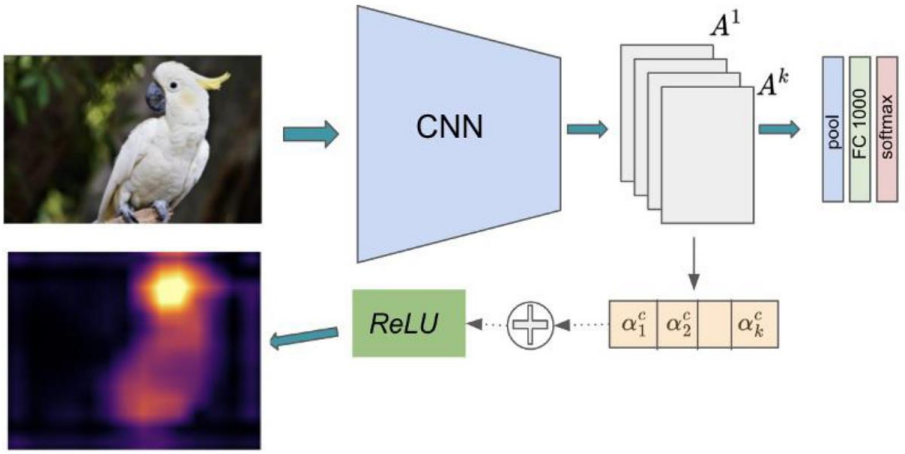
Hình 3. Kiến trúc GLFF

Các phương pháp gần đây đã đạt được những thành tựu đáng kể trong việc cải thiện độ chính xác và hiệu quả của hệ thống phân loại ảnh thật và ảnh do AI sinh ra. Tuy nhiên, đây vẫn là một bài toán khó và cần tiếp tục được nghiên cứu để đạt được kết quả tốt hơn.

2.4. Các phương pháp giải thích kết quả mô hình

2.4.1. Grad-CAM

Grad-CAM là một trong những kỹ thuật giải thích đầu tiên được phát triển cho các mô hình xử lý ảnh và có thể được áp dụng lên các bất kỳ mạng CNN nào. Nó tổng quát hóa kỹ thuật CAM thứ mà chỉ có thể dùng được cho một số kiến trúc nhất định. Grad-CAM hoạt động dựa trên thông tin gradient đi qua lớp tích chập cuối (hoặc bất kỳ) của mạng. Kết quả cho ra một heat map đánh dấu vùng của hình ảnh có ảnh hưởng cao nhất tới dự đoán của mô hình về lớp đã cho trước.



Hình 4. Sơ đồ hoạt động của phương pháp GRAD-CAM

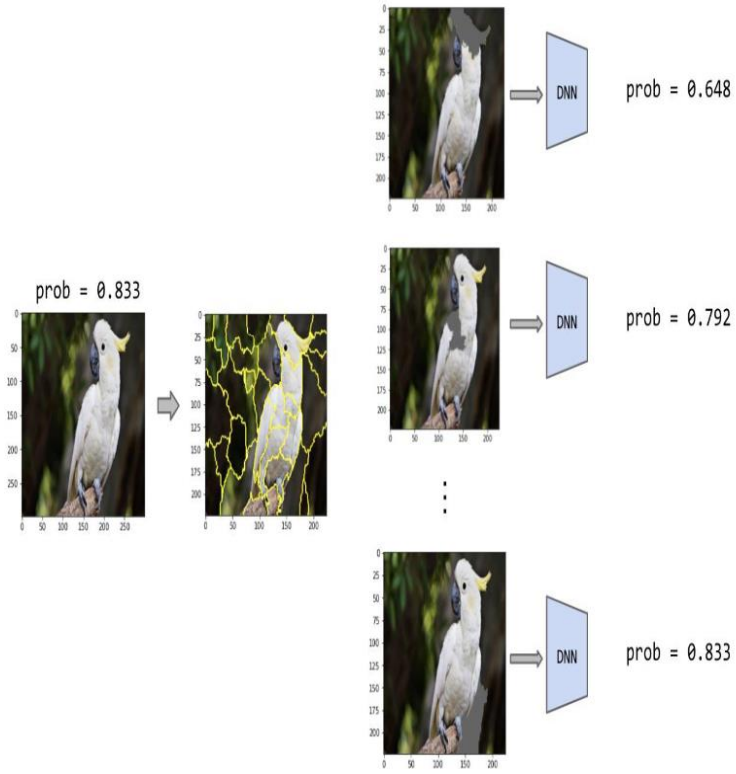
Grad-CAM là một kĩ thuật giải thích hậu xử lý và không yêu cầu bất kỳ thay đổi nào về kiến trúc hay huấn luyện [7]. Thay vào đó, Grad-CAM truy cập vào lớp tích chập bên trong của mô hình để xác định khu vực có ảnh hưởng cao nhất đến sự dự đoán của mô hình. Bởi vì Grad-CAM chỉ dựa vào các bước chuyển tiếp qua mô hình, không có lan truyền ngược nên nó cũng có hiệu quả về mặt tính toán.

Hình 5. Mô tả hoạt động của activation map trong GRAD-CAM

2.4.2. LIME

LIME là một trong những kỹ thuật giải thích phổ biến nhất. LIME được sử dụng sau khi đã huấn luyện xong mô hình. LIME có thể được dùng cho các mô hình regression và classification Về bản chất, LIME coi mô hình đã huấn luyện như một API, lấy các mẫu và tạo ra các giá trị dự đoán [7].

LIME giải thích bằng sự nhiễu loạn xảy ra ở cấp độ đặc trưng của đầu vào. Ở ảnh, các sự nhiễu loạn này được biểu diễn ở mức độ pixel và vùng pixel. Bằng cách này, những pixel và vùng pixel có ảnh hưởng cao nhất với sự dự đoán mô hình được đánh dấu rằng tăng hay giảm sự dự đoán của mô hình với đầu vào đã cho. Kết quả cuối cùng thể hiện các vùng có ảnh hưởng cao nhất đối với lớp phân loại được chọn.



Hình 6. Sơ đồ hoạt động phương pháp LIME

3. GIẢI PHÁP TRIỂN KHAI

3.1. Giải pháp về dữ liệu

3.1.1. Thu thập ảnh mặt người thật

Với dữ liệu ảnh khuôn mặt người thật, nhóm sử dụng bộ dữ liệu ***Flickr-Faces-HQ Dataset*** (FFHQ) [12]. FFHQ là bộ dữ liệu mặt người do NVIDIA công bố vào năm 2019, bao gồm khoảng 70.000 ảnh mặt người có kích thước 1024 x 1024 với chất lượng cao và đa dạng về quốc tịch, độ tuổi, nền ảnh, nhiều loại phụ kiện như mũ, kính mắt và kính râm trong ảnh.

Kể từ khi phát hành, bộ dữ liệu này đã trở thành bộ dữ liệu khuôn mặt được sử dụng rộng rãi nhất cho nhiều ứng dụng nghiên cứu và thương mại khác nhau, từ nhận dạng khuôn mặt đến nhận dạng giới tính và đặc biệt là dùng để huấn luyện các mô hình sinh ảnh mặt người giả.

Hình 7. Một số hình ảnh trong bộ FFHQ

3.1.2. Thu thập ảnh mặt người do AI tạo ra

Hiện nay, có khá nhiều nguồn ảnh do AI tạo ra. Sau quá trình tìm hiểu, các nguồn ảnh mặt người do AI tạo ra có chất lượng tốt mà nhóm có thể tiếp cận được bao gồm:

- *thispersondoesnotexist.com*: website này sẽ trả về một ảnh mặt người giả được sinh ra từ StyleGAN2 sau mỗi lần làm mới.
- *SFHQ phần 1* [11]: ảnh được tạo ra bằng cách đưa các tranh vẽ, tranh 3D vào mô hình StyleGAN2 để tăng độ chân thật.
- *SFHQ phần 2* [11]: lấy các ảnh 3D và các ảnh sinh ra từ Stable Diffusion 1.4 và tiến hành cải thiện độ chân thật với StyleGAN2.
- *SFHQ phần 3* [11]: chứa các ảnh được sinh trực tiếp từ StyleGAN2.
- *SFHQ phần 4* [11]: dùng Stable Diffusion 2.1 để chuyển văn bản thành hình ảnh và dùng StyleGAN để cải thiện chất lượng.
- *generated.photos*: website cung cấp nhiều hình ảnh mặt người giả được tạo ra từ StyleGAN.
- *Stable Diffusion*: được sinh bằng cách sử dụng prompt và mô hình Stable Diffusion XL 1.0.

Trong các nguồn ảnh trên, nhóm tiến hành tự thu thập dữ liệu từ các nguồn thispersondoesnotexist.com, generated.photos.Với Stable Diffusion, nhóm xây dựng một bộ prompt đa dạng và tiến hành sinh ảnh sử dụng mô hình Stable Diffusion XL 1.0. Đối với dữ liệu từ bộ SFHQ nhóm sử dụng bộ dữ liệu có sẵn trên Kaggle, với mỗi phần, nhóm tiến hành lấy ra 5000 ảnh để sử dụng.

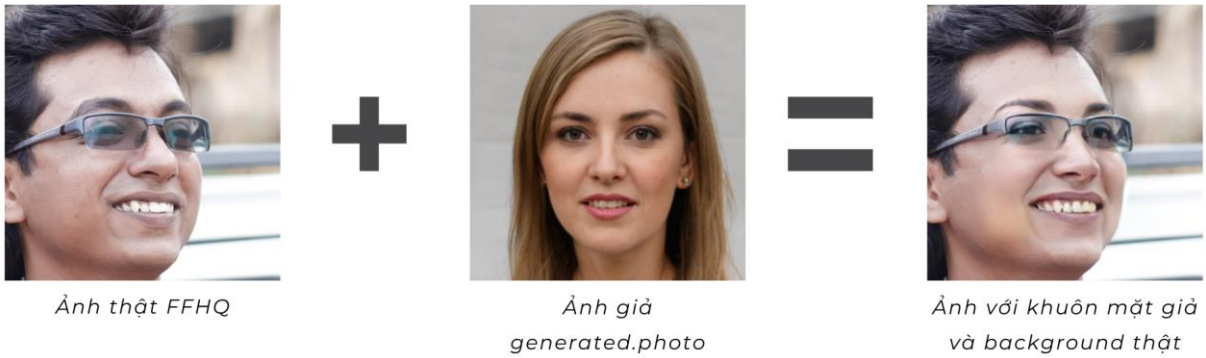
Dataset/ Data source	Số lượng ảnh thu thập/ sử dụng	Kích thước gốc của ảnh	Độ chân thật của hình ảnh
thispersondoesnotexist.com	20,000	1024 x 1024	Cao
SFHQ phần 1	5,000	1024 x 1024	Thấp
SFHQ phần 2	5,000	1024 x 1024	Thấp
SFHQ phần 3	5,000	1024 x 1024	Trung bình
SFHQ phần 4	5,000	1024 x 1024	Cao
generated.photos	10,000	256 x 256	Cao
Stable Diffusion	1,000	1024 x 1024	Trung bình

Bảng 1. Tổng quát thu thập dữ liệu

3.1.3. Tiền xử lý dữ liệu

Trong quá trình khảo sát, nhóm nhận thấy dữ liệu thu thập từ website generated.photos có nền chủ yếu là các màu đơn sắc như trắng, xám, nâu,... Điều này có thể dẫn đến mô hình bị thiên kiến với các ảnh có nền đơn sắc. Bên cạnh đó, tuy có chất lượng sinh ảnh giả cao nhưng độ phân giải của ảnh thu tập được trên website này khá thấp (256x256). Để giải quyết vấn đề trên, nhóm tiến hành xử lý bằng cách sử dụng ảnh thật từ bộ FFHQ và tiến hành swap/blend khuôn mặt giả từ bộ ảnh generated.photos vào các ảnh thật này. Lúc này ta có một ảnh mới với background ảnh là thật và khuôn mặt giả mạo.

Theo đánh giá chủ quan, phương pháp này có thể giúp đa dạng hóa thêm bộ dữ liệu và khắc phục vấn đề bias của mô hình như đã đề cập cũng như cải thiện điểm yếu của bộ dữ liệu generated.photos là độ phân giải thấp. Trong nghiên cứu này, nhóm sử dụng API Face Dancer của tác giả Felix Rosberg [2] vì API này có tốc độ xử lý khá nhanh, chất lượng swap cao và miễn phí.



Hình 8. Mô tả quá trình xử lý ảnh trong tập generated.photos

Trước khi sử dụng cho quá trình huấn luyện, nhóm tiến hành một số bước tiền xử lý dữ liệu sau:

- Toàn bộ dataset sẽ được resize về kích thước chung 512x512 và nén về định dạng JPEG nhằm giảm kích thước dữ liệu cho phù hợp với tài nguyên huấn luyện hiện có của nhóm.
- Chuẩn hóa dữ liệu: đưa giá trị các pixel về phạm vi [0, 1]. (Với một số kiến trúc CNN mà nhóm thử nghiệm như kiến trúc MobileNet, EfficientNet,.. có thể bỏ qua bước này vì trong kiến trúc đã có sẵn lớp chuẩn hóa dữ liệu).
- Tiếp theo, nhóm tiến hành làm giàu dữ liệu nhằm mô phỏng các thay đổi trên ảnh mà có thể xảy ra trong thực tế như xoay, lật, tăng giảm độ sáng, tương phản,v.v... (Chỉ thực hiện tăng cường dữ liệu trên tập huấn luyện).

Sau quá trình tiền xử lý, dữ liệu sẽ được chia thành 3 bộ train/test/validate với tỷ lệ chia 6/2/2 và dữ liệu ảnh giả sẽ được chia đều cho các nguồn ảnh giả khác nhau.

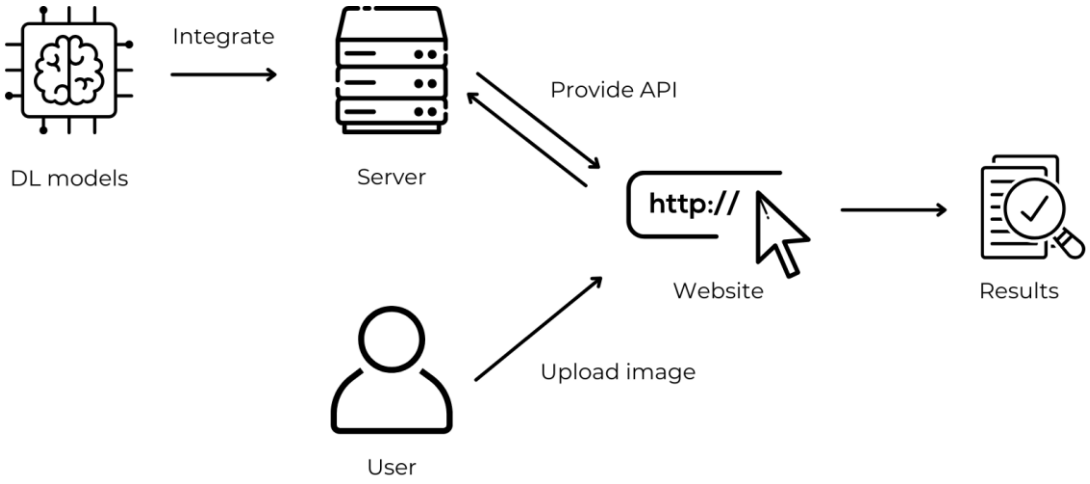
Hình 9. Quy trình tiền xử lý dữ liệu

3.2. Giải pháp về hệ thống

3.2.1. Xây dựng cấu trúc hệ thống và dịch vụ

Hệ thống cung cấp dịch vụ bao gồm ba thành phần chính:

- Server: được xây dựng dựa trên Framework FastAPI.
- Website clients: được thiết kế sử dụng HTML, CSS và JavaScript.
- Mô hình học sâu phân loại ảnh do AI tạo ra.



Hình 10. Sơ đồ tổng thể của hệ thống

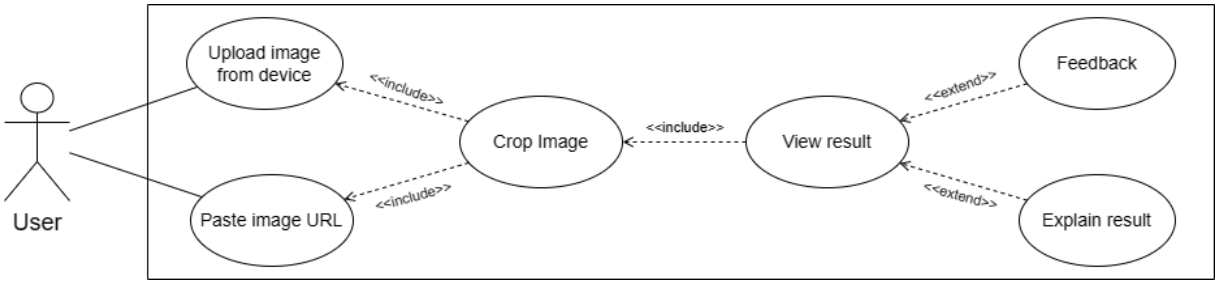
Cụ thể:

- Nhóm sử dụng FastAPI để xây dựng Server hệ thống, cho phép giao tiếp với người dùng thông qua Website.
- Mô hình học sâu được tích hợp trong Server, cung cấp API nhận dạng hình ảnh đến website dịch vụ.
- Khi người dùng chọn ảnh muốn kiểm tra, Website sẽ gửi ảnh cho Server xử lý. Sau đó, Server sẽ gửi phản hồi lại cho Website hiển thị kết quả.

Hình 11. FastAPI

- Về Website clients, nhóm tiến hành xây dựng một Website đơn giản với HTML, CSS và Javascript để tương tác với người dùng cũng như giao tiếp với Server.

3.2.2. Sơ đồ usecase hệ thống



Hình 12. Sơ đồ usecase hệ thống

Chức năng	Mô tả
Upload image from device	Người dùng upload ảnh từ thiết bị lên hệ thống để nhận dạng.
Paste image URL	Người dùng có thể dán URL dẫn đến hình ảnh cần nhận dạng.
Crop image	Người dùng có thể phóng to, thu nhỏ và cắt ảnh.
View result	Người dùng xem kết quả ảnh đó có phải là do AI tạo ra hay không.
Feedback	Người dùng có thể đánh giá kết quả hệ thống trả về là đúng hay sai.
Explain result	Người dùng xem giải thích kết quả dự đoán của hệ thống.

Bảng 2. Phân tích chức năng hệ thống

3.2.3. Tính năng phát hiện khuôn mặt người

Vì mô hình phân loại chỉ hỗ trợ nhận diện ảnh có đúng một mặt người. Do đó, nhóm tiến hành tích hợp thêm kĩ thuật **Haar Cascade Opencv Python** nhằm phát hiện khuôn mặt và loại bỏ đi hai trường hợp: ảnh không có mặt người và ảnh có nhiều hơn một mặt người. Ý tưởng chính của kỹ thuật trên:

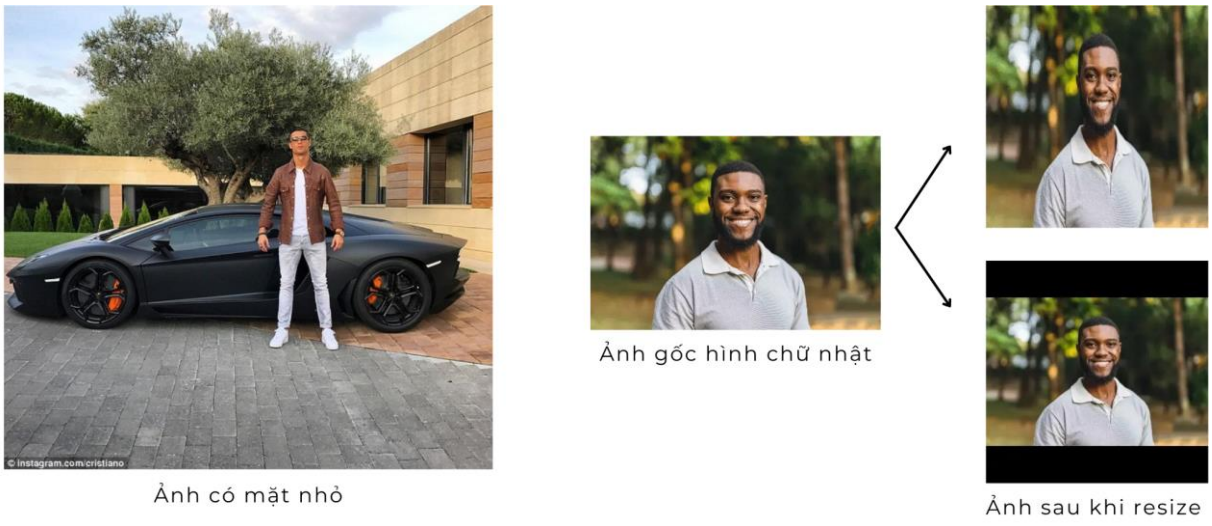
- Bước 1: Chia ảnh thành các ô nhỏ để tăng cường khả năng nhận diện.
- Bước 2: Mỗi ô nhỏ trên ảnh được kiểm tra bởi một chuỗi các bộ dò Haar.
- Bước 3: Nếu một ô nhỏ được xác định là có khả năng chứa khuôn mặt, nó được chọn làm kết quả tiềm năng và được gửi đến một bộ tích lũy.

- Bước 4: Hệ thống chuyển đến ô nhỏ tiếp theo trên ảnh và bắt đầu lại quá trình từ bước 2. Quá trình này được lặp lại cho toàn bộ ảnh.

Hình 13. Mô tả tính năng phát hiện khuôn mặt người

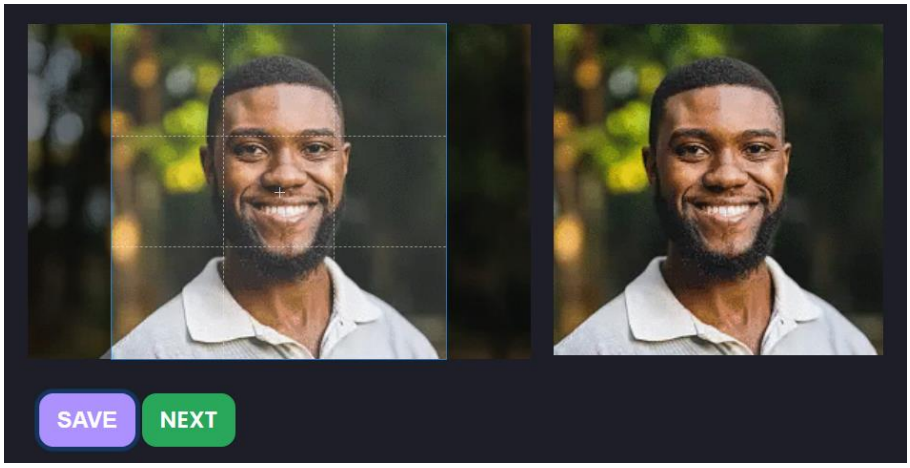
3.2.4. Tính năng cắt ảnh

Qua quá trình khảo sát thực tế, có hai vấn đề về ảnh đầu vào gây ảnh hưởng tới kết quả dự đoán của mô hình như ảnh có khuôn mặt nhỏ hơn nhiều so với kích thước ảnh tổng thể hoặc ảnh hình chữ nhật khi tiến hành thay đổi kích thước trước khi đưa vào mô hình sẽ làm ảnh sẽ bị biến dạng đáng kể hoặc khiến ảnh chứa các dữ liệu không mong muốn.



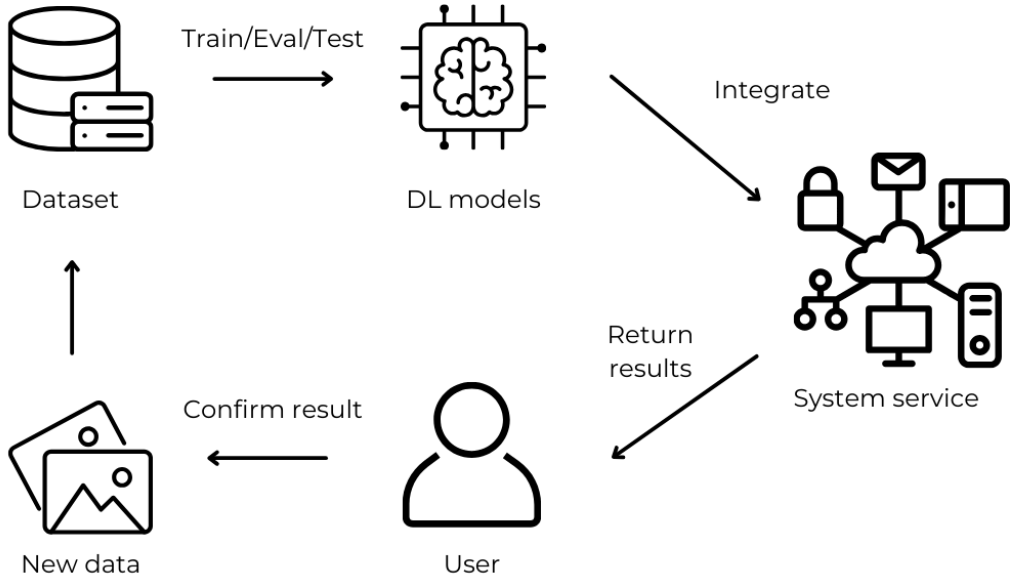
Hình 14. Các vấn đề của ảnh đầu vào

Giải pháp nhóm đưa ra cho hai vấn đề này đó là tính năng crop image (cắt ảnh) trên giao diện người dùng. Trang này sẽ cho phép người dùng phóng to, thu nhỏ và cắt ảnh gốc thành ảnh mới có kích thước vuông. Với phương pháp này, người dùng có thể dễ dàng điều chỉnh vùng ảnh muốn phân loại mà không ảnh hưởng đến kết quả dự đoán của mô hình.



Hình 15. Mô tả tính năng cắt ảnh

3.2.5. Tính năng lưu lại feedback của người dùng



Hình 16. Sơ đồ quá trình cập nhật lại hệ thống

Trong quá trình hoạt động, hệ thống có thể bị lỗi thời khi các mô hình sinh ảnh mới xuất hiện. Do đó, hệ thống cần được cập nhật thường xuyên với dữ liệu mới. Nguồn dữ liệu mới này có thể đến từ 2 nguồn chính:

- Dữ liệu feedback từ người dùng: dựa vào feedback của người dùng đối với kết quả trả về từ hệ thống. Chúng ta có thể sử dụng nguồn dữ liệu này để cập nhật.
- Dữ liệu mới do nhóm tiếp tục thu thập và update cho mô hình.

3.3. Giải pháp phân biệt ảnh mặt người do AI tạo ra sử dụng deep learning

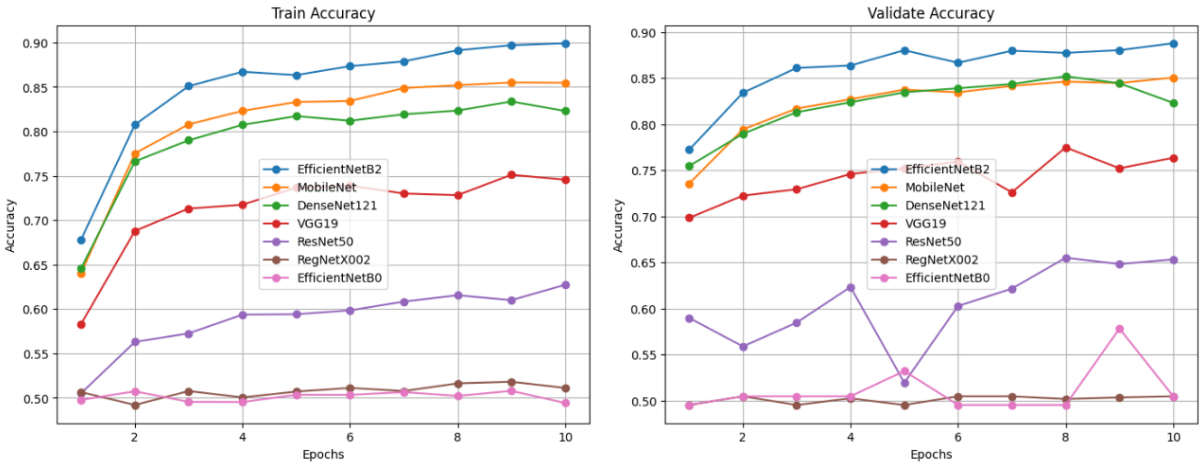
3.3.1. Transfer Learning và Ensemble Learning

Transfer learning là kỹ thuật huấn luyện tái sử dụng một mô hình đã được huấn luyện như điểm khởi đầu cho một mô hình mới với một nhiệm vụ khác. Ứng dụng transfer learning có thể giúp cải thiện độ chính xác của mô hình và đồng thời giảm thiểu thời gian huấn luyện.

Ensemble learning là phương pháp giúp tăng độ chính xác trên tập dữ liệu bằng cách kết hợp một số mô hình với nhau. Có nhiều kỹ thuật ensemble learning khác nhau và trong nghiên cứu này, nhóm tiến hành thử nghiệm hai kỹ thuật Concatenation và Average Ensemble.

3.3.2. Các kiến trúc CNN phổ biến cho bài toán Image Classification

Nhóm tiến hành lựa chọn các mô hình CNN phổ biến với bài toán phân loại ảnh bao gồm: VGG19, MobileNetV3, EfficientNetB0, EfficientNetB2, ResNet50, DenseNet121, RegNetX002 và thử nghiệm với kỹ thuật Transfer Learning. Nhóm sử dụng một tập dữ liệu nhỏ để thử nghiệm bao gồm 7000 ảnh mỗi nhãn ảnh thật và ảnh do AI tạo ra với 10 epochs.



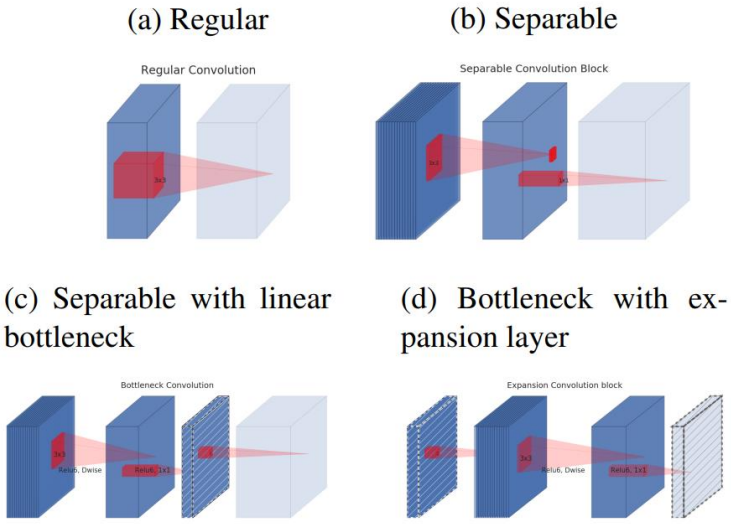
Hình 17. Kết quả thử nghiệm mô hình

Việc Transfer Learning với các kiến trúc CNN bước đầu cho kết quả khả quan. Nhóm tiếp tục lựa chọn hai mô hình có kết quả tốt nhất ở đây là ***MobileNetV3*** và ***EfficientNetB2*** để tiếp tục phát triển cho nghiên cứu.

3.3.3. Xây dựng mô hình hệ thống

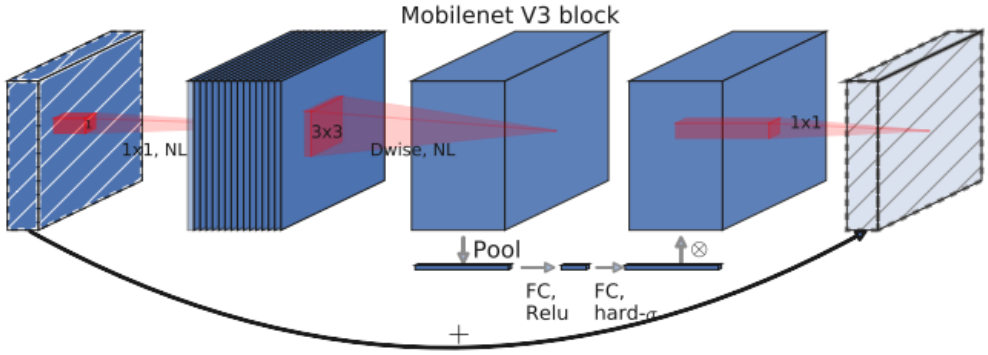
❖ **Kiến trúc mô hình MobileNetV3**

MobileNet là mô hình CNN được thiết kế với mục đích trở nên gọn nhẹ để ứng dụng vào các thiết bị di động và thiết bị nhúng.



Hình 18. Sự phát triển của Separable Convolution [4]

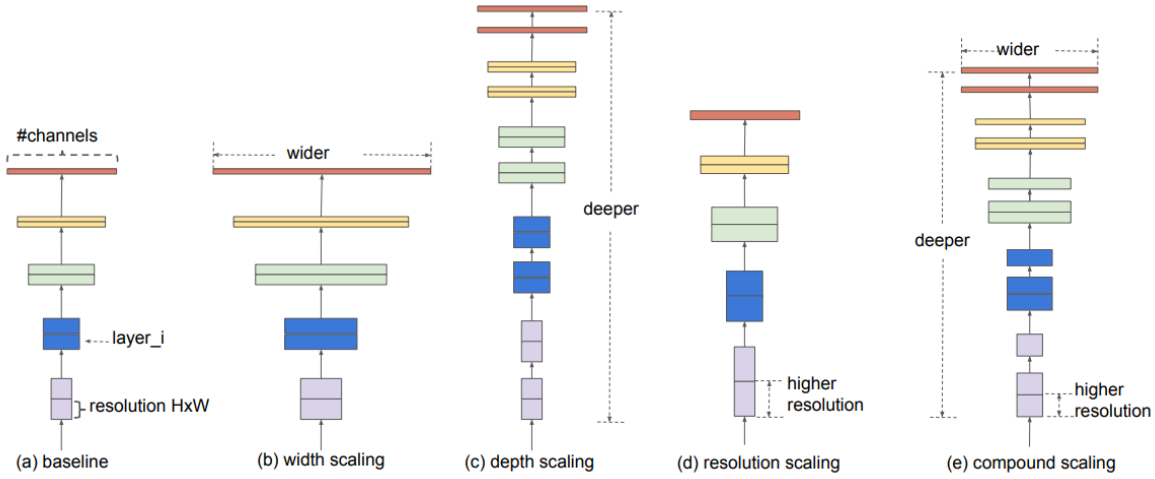
Depthwise separable convolution ở MobileNet đã giảm khối lượng tính toán và giảm số lượng tham số [3], đồng thời có thể thực hiện trích xuất đặc trưng một cách riêng biệt trên từng channel. MobileNetV2 đề xuất thêm Linear Bottlenecks giúp giảm kích thước input và Inverted Residual Block giúp tăng độ chính xác của mô hình mà không cần đến chi phí lớn [4]. MobileNetV3 tiếp tục cải tiến bằng việc dùng Squeeze and Excite nhằm tăng lượng thông tin giữa các kênh [5].



Hình 19. Kiến trúc MobileNet V3 [5]

❖ **Kiến trúc mô hình EfficientNet**

EfficientNet xoay quanh khái niệm thu phóng mô hình (Model Scaling) với các chiều sâu, rộng, phân giải.



Hình 20. Model Scaling [6]

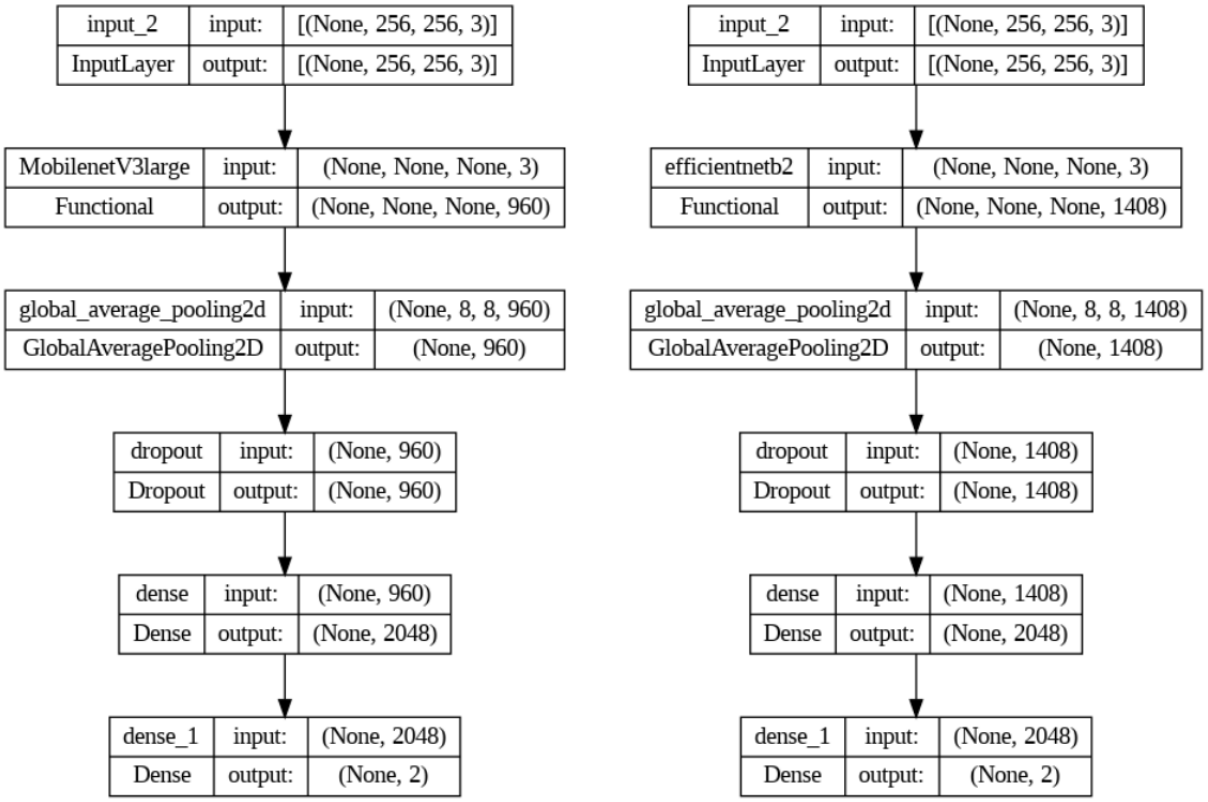
Mỗi việc thu phóng riêng lẻ có thể không tối ưu với đa số các bài toán mà còn mang hiệu ứng ngược. Vì vậy để đạt được độ chính xác và hiệu quả tốt hơn, điều quan trọng là phải cân bằng tất cả các kích thước của chiều rộng, chiều sâu và độ phân giải mạng trong quá trình thu phóng quy mô mạng nơ ron tích chập [6]. Từ đó các mô hình phiên bản của EfficientNet được phát triển cùng với các kỹ thuật được lấy cảm hứng từ MobileNet và ResNet.

❖ **Các mô hình đề xuất**

Sử dụng backbone là kiến trúc ***MobileNetV3*** và ***EfficientNetB2***, nhóm tiến hành xây dựng 4 mô hình sử dụng hai kỹ thuật chính là Transfer Learning và Ensemble Learning.

Với kỹ thuật Transfer Learning, nhóm sử dụng mô hình pretrained của hai kiến trúc trên làm bộ trích xuất đặc trưng. Sau đó nhóm tiến hành thêm các lớp Global Average Pool, Dropout, Fully Connected và Softmax vào sau bộ trích xuất đặc trưng. Từ đó nhóm xây dựng được hai mô hình:

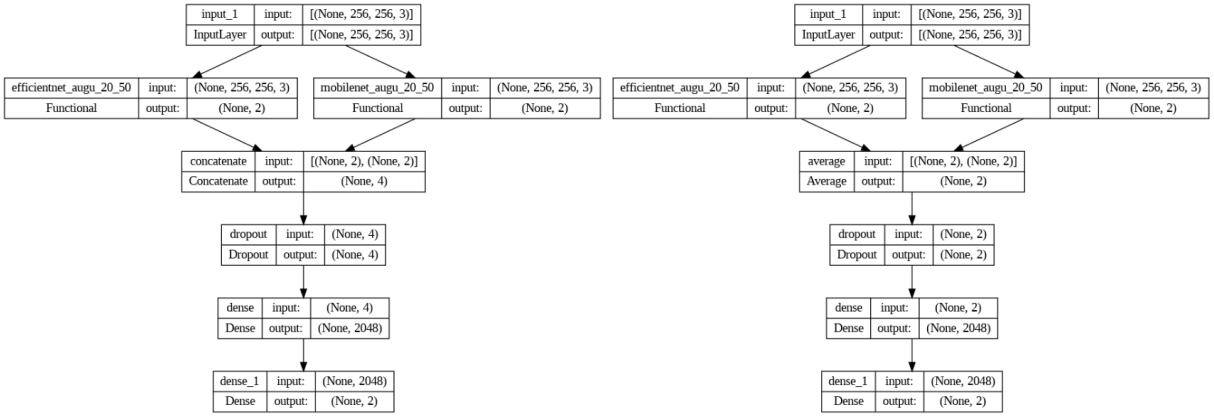
- ***Mô hình 1:*** sử dụng kỹ thuật Transfer Learning với backbone MobileNetV3.
- ***Mô hình 2:*** sử dụng kỹ thuật Transfer Learning với backbone EfficientNetB2.



Hình 21. Cấu trúc mô hình 1 (trái) và mô hình 2 (phải)

Sau khi huấn luyện hai mô hình 1 và mô hình 2, nhóm sử dụng thêm hai kỹ thuật Concatenation Ensemble và Average Ensemble để tiến hành kết hợp hai mô hình trên. Từ đó, nhóm xây dựng thêm hai mô hình.

- ***Mô hình 3:*** sử dụng kỹ thuật Concatenation Ensemble để kết hợp hai mô hình 1 và mô hình 2.
- ***Mô hình 4:*** sử dụng kỹ thuật Average Ensemble để kết hợp hai mô hình 1 và mô hình 2.



Hình 22. Cấu trúc mô hình 3 (trái) và mô hình 4 (phải)

3.3.4. Huấn luyện mô hình

Hai mô hình 1 và mô hình 2 được huấn luyện với các thông số sau:

- Kích thước ảnh đầu vào: 256x256 với batch size được sử dụng là 256. Huấn luyện trong 50 epochs và finetuning trong 10 epochs. Learning rate khởi tạo cho quá trình huấn luyện là 0.01 và cho quá trình finetuning là 0.001. Trong quá trình training, giảm learning rate sau mỗi 5 epochs, và sau mỗi 3 epochs trong quá trình finetuning.
- Optimizer được sử dụng là Adam optimizer. Hàm loss được sử dụng là Binary Crossentropy. Dừng huấn luyện khi mô hình không cải thiện trong 10 epochs, dừng finetune khi mô hình không cải thiện trong 6 epochs.
- Với mô hình 1 nhóm tiến hành finetune trên 63 layers của backbone MobileNetV3, với mô hình 2 nhóm tiến hành finetune trên 40 layers của backbone EfficientNetB2.

Hai mô hình 3 và mô hình 4 được huấn luyện chủ yếu để cập nhật các tham số của các lớp phân loại ở cuối của mô hình, do đó nhóm chỉ huấn luyện mô hình trong 10 epochs với learning rate khởi tạo là 0.001. Giảm learning rate sau mỗi 2 epochs và dừng huấn luyện khi mô hình không cải thiện trong 4 epochs.

Cả 4 mô hình trên đều được huấn luyện trên cấu hình chung như sau:GPU NVIDIA V100 - 16GB vRAM; CPU Intel(R) Xeon(R) - 2.00GHz; RAM: 51 GB.

4. THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

4.1. Kết quả thu thập và xử lý dữ liệu

Tổng kết, nhóm đã xây dựng được bộ dữ liệu bao gồm 120,959 ảnh thuộc về hai lớp là ảnh thật (70,000 ảnh) và ảnh do AI tạo ra (50,959) ảnh. Bộ dữ liệu trên được chia thành 3 tập train/test/validation theo tỷ lệ 6/2/2, các lớp ảnh giả được chia đều cho mỗi lớp train/test/validate.

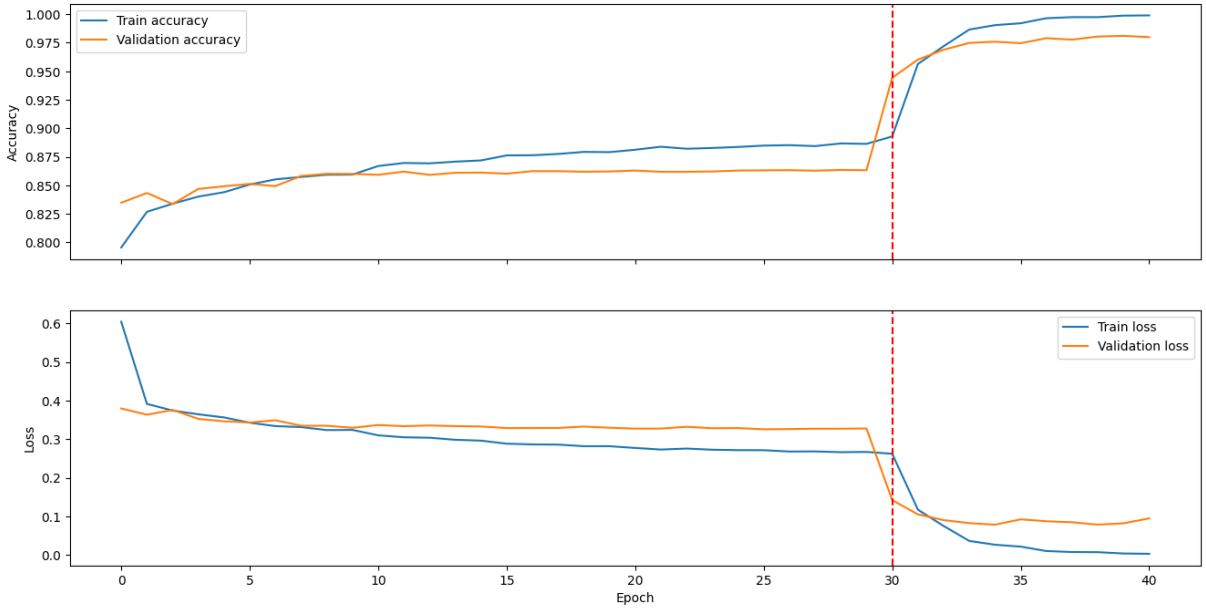
Lớp	Bộ dữ liệu	Số lượng ảnh
Ảnh thật	FFHQ	70,000
Ảnh do AI sinh ra	thispersondoesnotexist.com	20,499
	SFHQ	20,000
	Swap face	9,429
	Stable Diffusion	1,031

Bảng 3. Số lượng dữ liệu đã thu thập

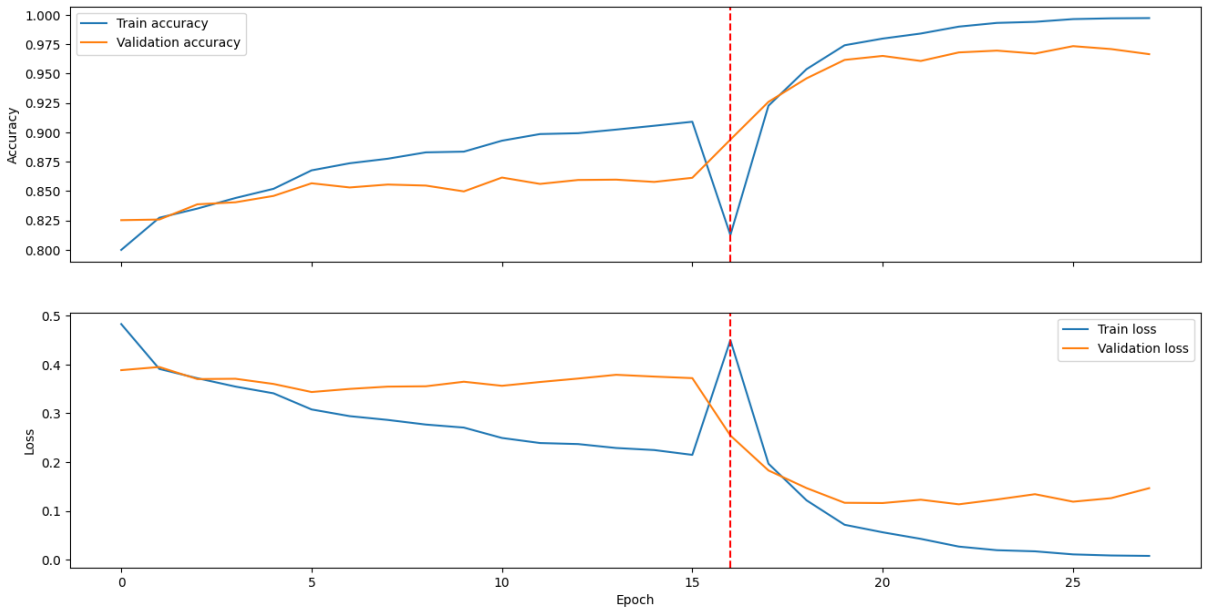
4.2. Kết quả huấn luyện và kiểm thử mô hình

4.2.1. Kết quả huấn luyện mô hình

Sau đây là kết quả huấn luyện của hai mô hình 1 và mô hình 2. Đối với hai mô hình 3 và mô hình 4, nhóm sẽ không trình bày kết quả huấn luyện ở đây vì quá trình huấn luyện của hai kỹ thuật này chỉ chủ yếu để cập nhật tham số của hai lớp phân loại cuối cùng.



Hình 23. Kết quả huấn luyện mô hình 1



Hình 24. Kết quả huấn luyện mô hình 2

4.2.2. Kết quả kiểm thử mô hình

Nhóm tiến hành đánh giá mô hình trên tập kiểm thử dựa trên các metrics sau: accuracy, precision, F1 và recall.

Mô hình	Class	Precision	Recall	F1	Accuracy (%)
Mô hình 1	Real	**0.97**	**0.99**	**0.98**	**97.91**
	AI	**0.99**	**0.96**	**0.97**	
Mô hình 2	Real	0.98	0.96	0.97	96.58
	AI	0.95	0.97	0.96	
Mô hình 3	Real	0.97	0.99	0.98	97.86
	AI	0.99	0.96	0.97	
Mô hình 4	Real	0.97	0.99	0.98	97.85
	AI	0.99	0.96	0.97	

Bảng 4. Kết quả kiểm thử mô hình

Mô hình	Kích thước mô hình (MB)	Inference time trên tập kiểm thử (s)
Mô hình 1	**50.3**	**15**
Mô hình 2	90.8	28
Mô hình 3	60.6	36
Mô hình 4	60.6	36

Bảng 5. Đánh giá kích thước và thời gian chạy trên tập kiểm thử

4.3. Kết quả xây dựng và kiểm thử hệ thống dịch vụ

4.3.1. Kết quả xây dựng hệ thống dịch vụ

❖ **Cấu hình cài đặt Server**

- Host: 127.0.0.1
- Port: 8000
- Worker: 4
- Thông tin về định tuyến trong server được liệt kê trong bảng sau.

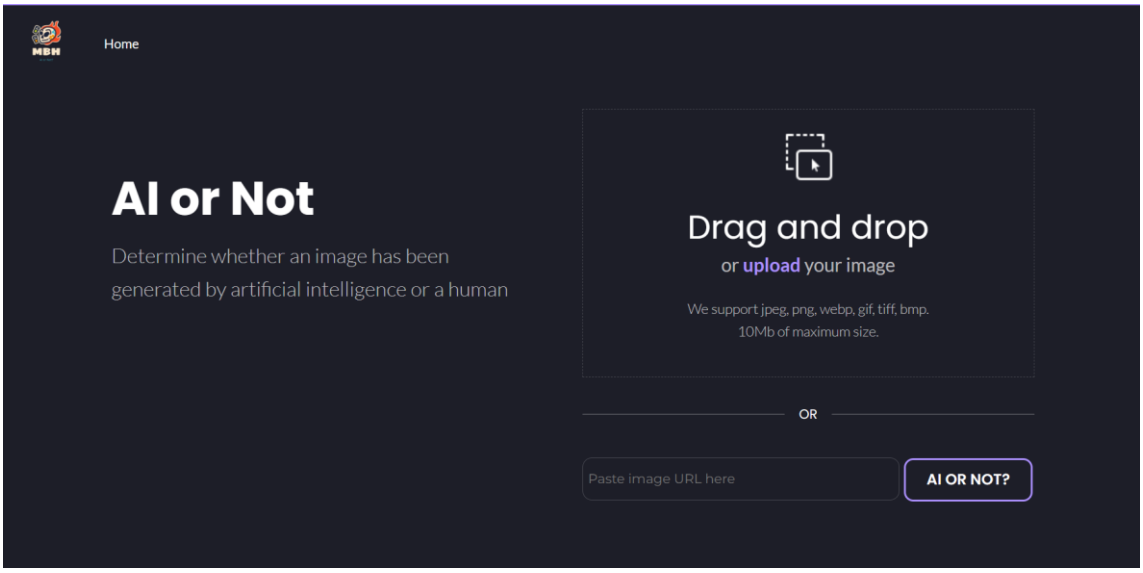
Endpoint	Phương thức	Tóm tắt	Tham số	Phản hồi
upload_and_process	POST	Lưu ảnh, xử lý và trả về kết quả dự đoán.	File ảnh	200: Phản hồi thành công, trả về kết quả dự đoán.

explain	GET	Nhận ảnh và giải thích kết quả dự đoán.	File ảnh	200: Phản hồi thành công, trả về ảnh giải thích kết quả dự đoán.
submit_feedback	POST	Nhận kết quả feedback của người dùng và lưu lại vào database.	Dict (0 hoặc 1)	200: Phản hồi thành công, trả về thông báo feedback thành công.

Bảng 6. Bảng thông tin định tuyến các endpoints

❖ **Trang chủ**

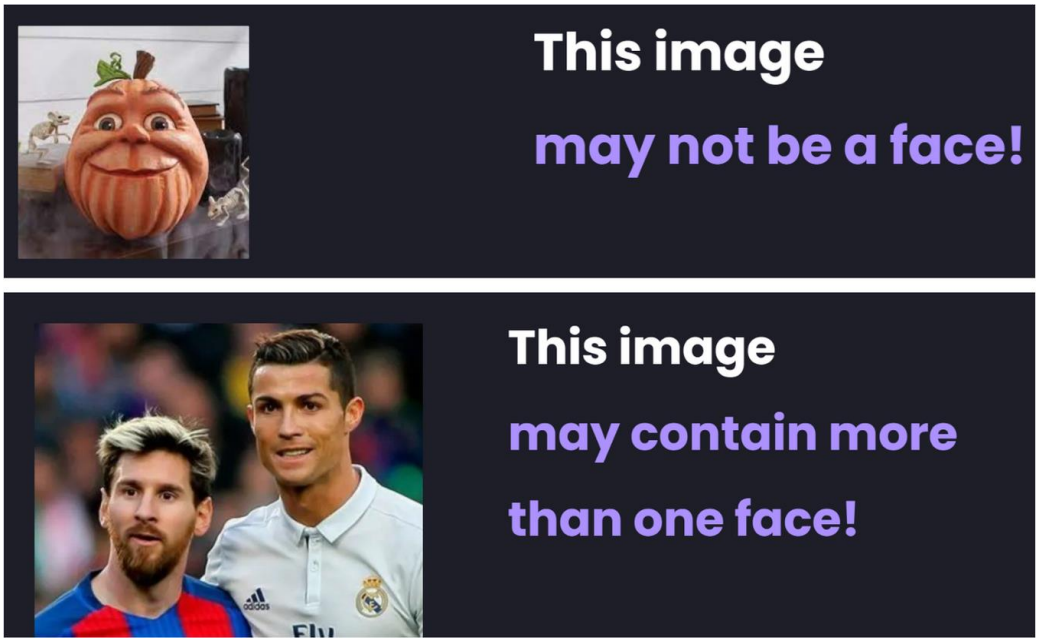
Trang chủ của giao diện hệ thống là nơi người dùng có thể tải ảnh từ thiết bị hoặc dán đường dẫn đến ảnh mà người dùng muốn nhận diện.



Hình 25. Giao diện trang chủ

❖ **Tính năng kiểm tra hình ảnh có chứa đúng một khuôn mặt hay không**

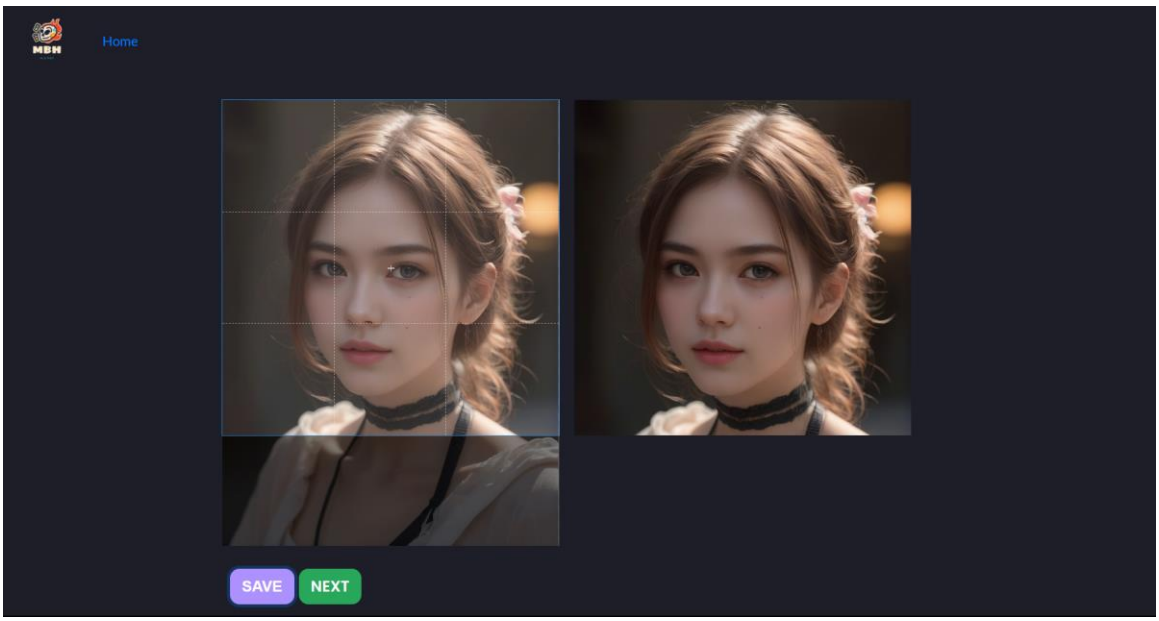
Sau khi người dùng tải ảnh lên hệ thống hoặc dán URL. Tính năng này giúp hệ thống phát hiện ảnh không chứa mặt người hoặc chứa nhiều hơn một mặt người.



Hình 26. Giao diện tính năng nhận diện khuôn mặt

❖ **Tính năng cắt ảnh**

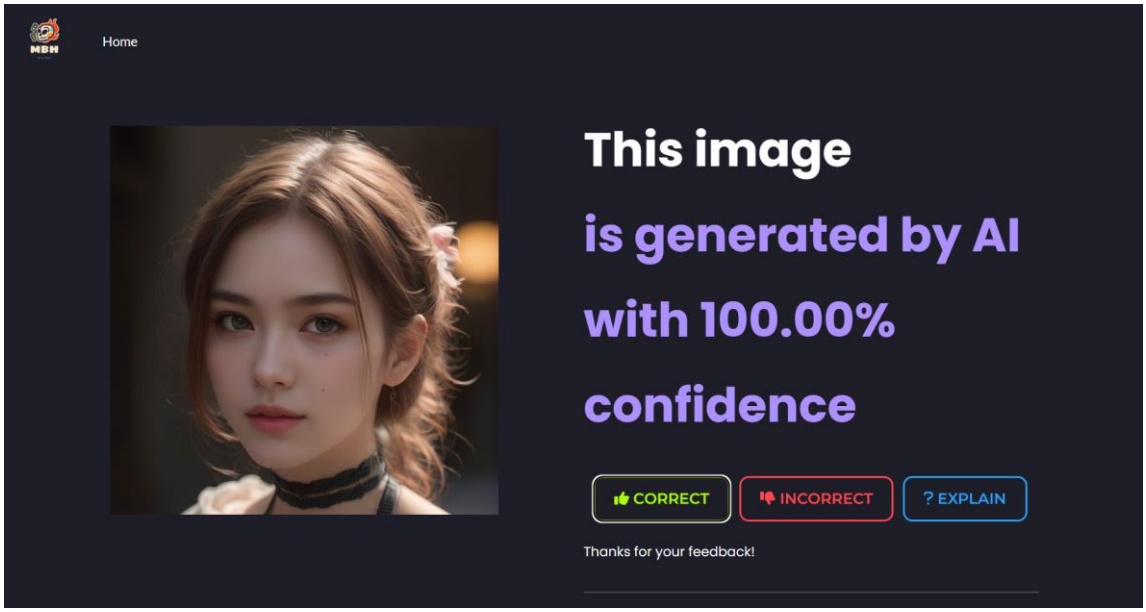
Tính năng này giúp người dùng có thể phóng to, thu nhỏ hình ảnh tùy thích và cắt phần ảnh muốn chọn để tiến hành nhận diện.



Hình 27. Giao diện tính năng cắt ảnh

❖ **Tính năng xem và phản hồi kết quả**

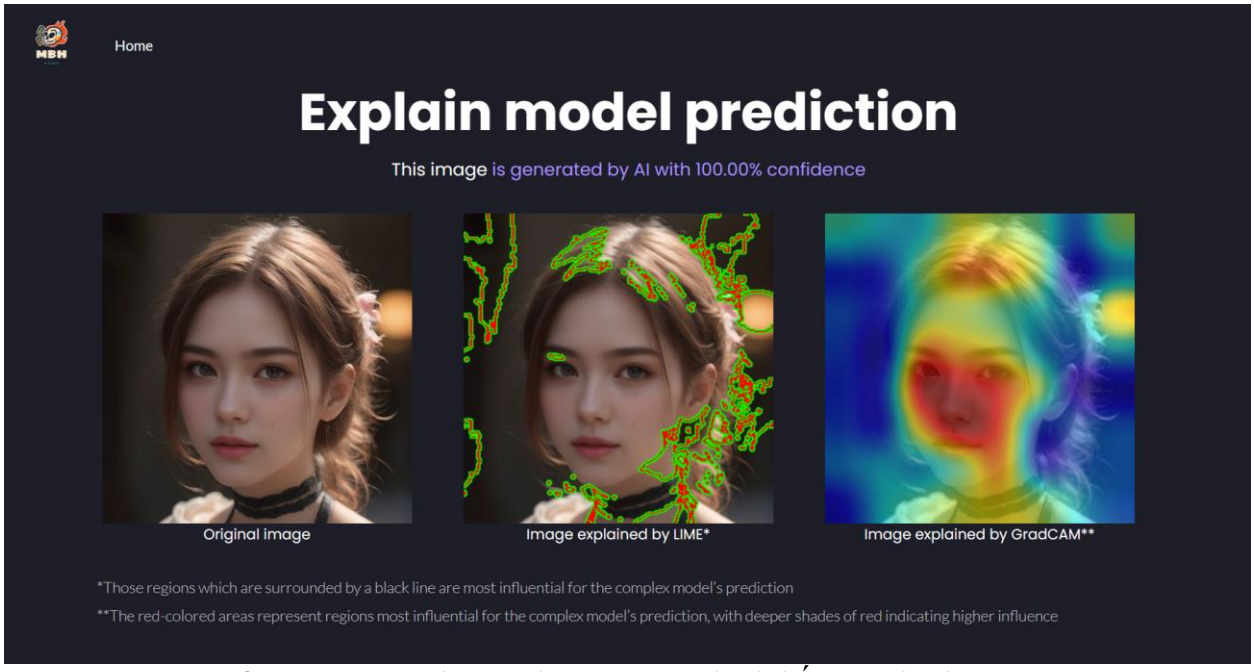
Sau khi đã cắt vùng ảnh người dùng mong muốn, hệ thống truyền ảnh vào mô hình học sâu để đưa ra dự đoán và trả về cho website. Người dùng có thể để lại feedback về kết quả dự đoán của hệ thống thông qua hai nút “Correct” và “Incorrect”



Hình 28. Giao diện tính năng xem và phản hồi kết quả

❖ **Giải thích kết quả dự đoán**

Khi người dùng bấm vào nút “Explain” hệ thống sẽ tiến hành giải thích kết quả dự đoán của mô hình sử dụng hai phương pháp LIME và Grad - CAM. Ở LIME các vùng được bọc bởi các đường tô đậm là những vùng có ảnh hưởng cao nhất tới dự đoán của mô hình. Ở Grad-CAM các vùng có ảnh hưởng tới mô hình được đánh dấu đỏ, sắc đỏ tăng dần theo mức độ ảnh hưởng tới dự đoán của mô hình.



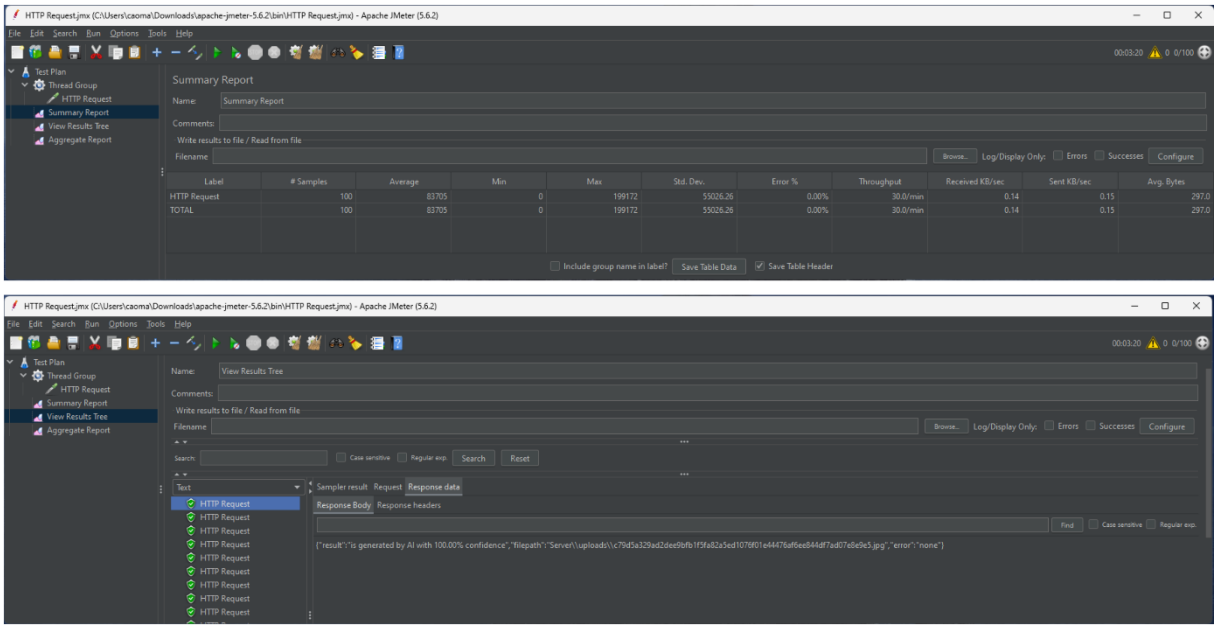
Hình 29. Giao diện tính năng giải thích kết quả dự đoán

4.3.2. Kết quả kiểm thử hệ thống dịch vụ

Nhóm tiến hành đánh giá thời gian xử lý của hệ thống trong thực tế với cấu hình server: CPU Intel(R) Core(TM) i3-1005G1 - 1.20GHz và RAM 4.00 GB.

❖ **Kiểm thử Server với Apache Jmeter**

Nhóm tiến hành kiểm thử khả năng đáp ứng của server với số lượng request threads là 100, cùng thực thi cùng một lúc, với phương thức POST upload_and_process.



Hình 30. Kết quả kiểm thử hệ thống với Apache JMeter

❖ **Đánh giá thời gian kiểm tra một ảnh có hợp lệ (chứa đúng một mặt người) hay không.**

Nhóm tiến hành đánh giá hệ thống với 10 ảnh trong đó không chứa mặt người hoặc chứa nhiều hơn một mặt người. Thời gian kiểm tra trung bình là xấp xỉ **1.24s/ảnh.**

❖ **Đánh giá thời gian chạy thực tế của 4 mô hình:**

Nhóm tiến hành đánh giá thời gian chạy của 4 mô hình khi tích hợp vào hệ thống với 10 ảnh. Lưu ý, thời gian đưa ra dự đoán của mô hình đã bao gồm thời gian kiểm tra ảnh có hợp lệ (chứa đúng một mặt người) hay không.

Mô hình	**Thời gian trung bình để đưa kết quả dự đoán (s)**	**Thời gian trung bình để giải thích kết quả dự đoán (s)**
Mô hình 1	**7**	**74**
Mô hình 2	13	160
Mô hình 3	21	x
Mô hình 4	20	x

Bảng 7. Kết quả đánh giá thời gian chạy của hệ thống trong thực tế

Với hai mô hình 3 và mô hình 4, vì cấu trúc mô hình phức tạp nên chưa áp dụng được hai kỹ thuật Grad-CAM và LIME để giải thích kết quả nên chưa đánh giá được thời gian trung bình để giải thích kết quả dự đoán của hai mô hình này.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Qua nghiên cứu này, nhóm đã xây dựng thành công được một bộ dữ liệu đa dạng về ảnh mặt người do AI tạo ra cũng như đã tiến hành thử nghiệm và đề xuất việc sử dụng các mô hình học sâu CNN cùng các kỹ thuật Transfer Learning - Ensemble Learning để xây dựng mô hình phân loại ảnh mặt người do AI tạo ra. Bên cạnh đó, nhóm cũng đã tiến hành đánh giá hiệu quả của các mô hình khi áp dụng vào hệ thống thực tế, từ đó xây dựng thành công hệ thống dịch vụ với độ chính xác cao, tốc độ xử lý nhanh và có khả năng mở rộng về sau.

Mặc dù đã đạt được một số kết quả tích cực, tuy nhiên nghiên cứu của nhóm vẫn còn một số điểm hạn chế và khó khăn sau:

- Về mặt dữ liệu: tuy bộ dữ liệu của nhóm có chất lượng khá tốt và độ đa dạng cao, nhưng vì những hạn chế về tài nguyên tính toán dẫn đến việc nhóm phải giảm chất lượng và kích thước của bộ dữ liệu huấn luyện. Bên cạnh đó, việc thực hiện các kỹ thuật tăng cường dữ liệu cũng bị hạn chế vì lý do trên.
- Về mặt mô hình: tuy các mô hình đều đạt kết quả khá cao, nhưng cách tiếp cận hiện tại vẫn đang phụ thuộc khá nhiều về dữ liệu.

5.2. Hướng phát triển

Với những kết quả đạt được hiện tại và những hạn chế trong dự án, trong tương lai, nhóm dự định sẽ tiếp tục phát triển thêm với các hướng đi sau:

- Về mặt dữ liệu: tiếp tục thu thập và mở rộng bộ dữ liệu cũng như thử nghiệm thêm nhiều phương pháp làm giàu dữ liệu.
- Về mặt mô hình: tiếp tục nghiên cứu và phát triển thêm các mô hình khác để giảm sự phụ thuộc vào dữ liệu và tăng cường tính tổng quát hóa của hệ thống.
- Về mặt hệ thống dịch vụ: cải thiện thêm giao diện và trải nghiệm người dùng. Thử nghiệm thêm các phương pháp tăng tốc hệ thống cũng như tiến hành đánh giá hệ thống một cách tổng quát và kĩ càng hơn.

TÀI LIỆU THAM KHẢO

[1] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, Siwei Lyu, "GLFF: Global and Local Feature Fusion for AI-synthesized Image Detection" (2023), https://arxiv.org/pdf/2211.08615.pdf

[2] Rosberg, Felix and Aksoy, Eren Erdal and Alonso-Fernandez, Fernando and Englund, Cristofer, "FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping" (2023), https://arxiv.org/pdf/2210.10473.pdf

[3] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, Hartwig Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications" (2017), https://arxiv.org/pdf/1704.04861.pdf

[4] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen, " MobileNetV2: Inverted Residuals and Linear Bottlenecks" (2018), https://arxiv.org/pdf/1801.04381v4.pdf

[5] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, Hartwig Adam, "Searching for MobileNetV3" (2019), https://arxiv.org/pdf/1905.02244.pdf

[6] Mingxing Tan, Quoc V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" (2020), https://arxiv.org/pdf/1905.11946.pdf

[7] Michael Munn, David Pitman, "Explainable AI for Practitioners" (2022)

[8] Haodong Li, Bin Li, Shunquan Tan, Jiwu Huang, "Identification of Deep Network Generated Images Using Disparities in Color Components" (2020), https://arxiv.org/pdf/1808.07276.pdf

[9] Scott McCloskey, Michael Albright, "Detecting GAN-generated Imagery using Color Cues" (2018), https://arxiv.org/pdf/1812.08247.pdf

[10] Zeyang Sha, Zheng Li, Ning Yu, Yang Zhang, "DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models", (2023) https://arxiv.org/pdf/2210.06998.pdf

[11] David Beniaguev, "Synthetic Faces High Quality (SFHQ) dataset" (2022), https://github.com/SelfishGene/SFHQ-dataset

[12] Tero Karras, Janne Hellsten, "Flickr-Faces-HQ Dataset (FFHQ)", https://github.com/NVlabs/ffhq-dataset

[13] Tero Karras, Samuli Laine, Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks" (2019), https://arxiv.org/pdf/1812.04948.pdf