

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH



BÁO CÁO ĐỒ ÁN CHUYÊN NGÀNH
Chương trình Cử nhân Tài năng

TĂNG CƯỜNG ĐỘ TIN CẬY CỦA TRÌNH PHÁT
HIỆN HÌNH ẢNH DEEPFAKE BẰNG PHƯƠNG
PHÁP TRÒ CHƠI ĐỐI KHÁNG KẾT HỢP
KHẢ NĂNG GIẢI THÍCH

Giáo viên hướng dẫn:

ThS. PHAN THẾ DUY

TS. PHẠM VĂN HẬU

Nhóm sinh viên thực hiện:

PHAN NGUYỄN HỮU PHONG

22521090

CHÂU THẾ VĨ

22521653

TP. Hồ Chí Minh, tháng 07 năm 2025

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH



BÁO CÁO ĐỒ ÁN CHUYÊN NGÀNH
Chương trình Cử nhân Tài năng

TĂNG CƯỜNG ĐỘ TIN CẬY CỦA TRÌNH PHÁT
HIỆN HÌNH ẢNH DEEPFAKE BẰNG PHƯƠNG
PHÁP TRÒ CHƠI ĐỐI KHÁNG KẾT HỢP
KHẢ NĂNG GIẢI THÍCH

Giáo viên hướng dẫn:

ThS. PHAN THẾ DUY

TS. PHẠM VĂN HẬU

Nhóm sinh viên thực hiện:

PHAN NGUYỄN HỮU PHONG

22521090

CHÂU THẾ VĨ

22521653

TP. Hồ Chí Minh, tháng 07 năm 2025

LỜI CẢM ƠN

Trong suốt quá trình thực hiện và hoàn thành đồ án, nhóm đã nhận được rất nhiều sự quan tâm, định hướng, hỗ trợ và những lời động viên quý báu từ các thầy cô, gia đình và bạn bè. Nhân dịp này, nhóm xin bày tỏ lòng biết ơn sâu sắc đến tất cả những người đã đồng hành cùng nhóm trong hành trình nghiên cứu vừa qua.

Trước hết, nhóm xin gửi lời cảm ơn chân thành và sâu sắc đến thầy Phan Thế Duy và thầy Phạm Văn Hậu – giảng viên khoa Mạng Máy tính và Truyền thông, trường Đại học Công nghệ Thông tin – ĐHQG TP.HCM. Với sự tận tâm trong hướng dẫn, định hướng chuyên môn rõ ràng cùng sự hỗ trợ nhiệt tình trong suốt quá trình thực hiện đề tài, các thầy đã giúp nhóm hoàn thiện đồ án chuyên ngành một cách hiệu quả và bài bản.

Nhóm cũng xin chân thành cảm ơn các thầy cô giảng viên trường Đại học Công nghệ Thông tin – ĐHQG TP.HCM, đặc biệt là quý thầy cô khoa Khoa học Máy tính những người đã truyền đạt kiến thức, khơi dậy niềm đam mê học hỏi và tạo nền tảng vững chắc để nhóm có thể phát triển ý tưởng và hoàn thiện đề tài.

Bên cạnh đó, nhóm xin gửi lời cảm ơn sâu sắc đến gia đình và bạn bè – những người luôn bên cạnh động viên, chia sẻ và khích lệ tinh thần trong suốt thời gian nhóm thực hiện đồ án.

Một lần nữa, nhóm xin chân thành cảm ơn tất cả những sự giúp đỡ, đồng hành và yêu thương mà chúng em đã nhận được.

Phan Nguyễn Hữu Phong

Châu Thế Vĩ

MỤC LỤC

LỜI CẢM ƠN	i
DANH SÁCH HÌNH VẼ	iv
DANH SÁCH BẢNG	vi
MỞ ĐẦU	1
CHƯƠNG 1. TỔNG QUAN	2
1.1 Giới thiệu vấn đề	2
1.2 Các công trình nghiên cứu liên quan	5
1.3 Mục tiêu, đối tượng, phạm vi nghiên cứu	8
1.3.1 Mục tiêu nghiên cứu	8
1.3.2 Đối tượng nghiên cứu	8
1.3.3 Phạm vi nghiên cứu	9
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	10
2.1 GANs (Generative Adversarial Networks)	11
2.2 Deepfake Generation	12
2.2.1 Entire Face Synthesis (Tổng hợp toàn bộ khuôn mặt) .	12
2.2.2 Face Identity Swap (Hoán đổi danh tính khuôn mặt) .	13
2.2.3 Face Manipulation (Thao tác khuôn mặt)	13
2.3 Deepfake Detection	14
2.3.1 Spatial Domain Detection (Phát hiện trong miền không gian)	14
2.3.2 Frequency Domain Detection (Phát hiện trong miền tần số)	14
2.4 Adversarial Training	15
2.4.1 Nguyên lý hoạt động	15
2.4.2 Quy trình hoạt động	16
2.4.3 Mục tiêu và Kết quả	16

2.5 Explainability-Guided Approach	17
2.5.1 Khái niệm Explainability trong học sâu	17
2.5.2 Các phương pháp XAI phổ biến	18
2.5.3 Explainability-Guided Approach	25
2.5.4 Lợi ích của Explainability-Guided Approach	26
2.5.5 Ứng dụng trong phát hiện Deepfake	27
CHƯƠNG 3. PHƯƠNG PHÁP THỰC HIỆN	28
3.1 Model Overview	29
3.2 Detector with Multi-task Learning	29
3.2.1 Architecture	30
3.2.2 Detector Loss Function	30
3.3 Explainability-Guided Adversarial Refiner	31
3.3.1 Architecture and Refinement Process	31
3.3.2 Prediction Explainer (CAM)	32
3.3.3 Refiner Loss Function	32
3.4 Adversarial Training Framework	33
CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ	34
4.1 Bộ dữ liệu	34
4.2 Độ đo	35
4.3 Cài đặt thực nghiệm	35
4.3.1 So sánh với các Phương pháp khác	36
4.3.2 Nghiên cứu Tách biệt (Ablation Study)	37
CHƯƠNG 5. KẾT LUẬN	39
5.1 Điểm mạnh	39
5.2 Hạn chế	39
5.3 Hướng phát triển trong tương lai	40
TÀI LIỆU THAM KHẢO	41

DANH SÁCH HÌNH VẼ

1.1	Ảnh minh họa việc các đối tượng lừa đảo sử dụng công nghệ Deepfake nhằm thực hiện các hành vi mạo danh người khác cho các mục đích xấu.	2
1.2	Sơ đồ minh họa kiến trúc tổng quát của một mô hình CNN . .	3
1.3	Hình ảnh minh họa một robot đang cố gắng xây một cây cầu từ những mảnh ghép chỉ có hình dạng của một loại khuôn mặt duy nhất. Cây cầu bắt đầu sụp đổ ở điểm mà một mảnh ghép có hình dạng khuôn mặt mới, khác biệt không thể khớp vào. Điều này tượng trưng cho việc AI đã "quá khớp" vào dữ liệu huấn luyện và thất bại trong việc tổng quát hóa khi gặp phải một loại dữ liệu mới mà nó chưa từng được học.	4
1.4	Luồng hoạt động cơ bản của các mô hình deepfake detection. Các hình ảnh có thể được phân tích đặc trưng tần số trước hoặc được đưa trực tiếp vào mô hình DNN. Các này phụ thuộc quá mức vào các dấu vết đặc thù của một dataset nên có xu hướng bị overfitting.	5
1.5	"Where Deepfakes Gaze at?". Nghiên cứu này đã tận dụng những đặc trưng về các thành phần trên gương mặt giúp mô hình khai thác được nhiều khía cạnh hơn.	7
2.1	So sánh ảnh gốc và ảnh Deepfake. Các đặc điểm khuôn mặt đã bị chỉnh sửa tinh vi, gây khó khăn cho việc phát hiện.	10
2.2	Tổng quan về kiến trúc của mạng GAN (Generative Adversarial Network). Kiến trúc chính bao gồm một Generator và một Discriminator cạnh tranh nhau trong quá trình huấn luyện. . .	11
2.3	Hình ảnh các khuôn mặt được tạo sinh hoàn toàn bằng mô hình StyleGAN của NVIDIA. Dù trông giống người thật, tất cả đều là khuôn mặt giả (synthetic faces), không tồn tại trong thực tế.	12
2.4	Minh họa hoán đổi danh tính khuôn mặt sử dụng FSGAN. Khuôn mặt từ ảnh gốc (source image) được hoán đổi lên video mục tiêu (target video), kết quả nằm ở giữa.	13
2.5	Minh họa tính giải thích (Explainability) trong trí tuệ nhân tạo, nâng cao sự minh bạch và tin cậy của các mô hình học máy	17

2.6	Hình ảnh minh họa bản đồ nhiệt (CAM) từ lớp Global Average Pooling trong mạng CNN, thể hiện khả năng giải thích (explainability) của mô hình học sâu trong việc nhận diện và phân tích đặc trưng của hình ảnh, chẳng hạn như một chú chó Australian Terrier trên ghế xe có dây an toàn.	19
2.7	Sơ đồ Grad-CAM minh họa cách tạo bản đồ nhiệt từ gradient và lớp kích hoạt CNN, xác định khu vực quan trọng để phân loại hình ảnh (mèo và chó) trước softmax.	20
2.8	Sơ đồ minh họa phương pháp LIME, chuyển từ mô hình phân loại toàn cục phi tuyến phức tạp sang mô hình cục bộ tuyến tính đơn giản để giải thích.	22
2.9	Sơ đồ minh họa giải thích SHAP, thể hiện tầm quan trọng của các đặc trưng (như tuổi, giới tính, huyết áp, BMI) trong việc đóng góp vào dự đoán và đánh giá mô hình học máy.	23
2.10	Sơ đồ minh họa quy trình huấn luyện mô hình học sâu với Explainability-Guided Approach, tích hợp Explainer (ví dụ bản đồ giải thích như CAM) để định hướng tập trung vào các đặc trưng khuôn mặt quan trọng trong phát hiện Deepfake.	25
3.1	Sơ đồ tổng quan kiến trúc Trò chơi Đối kháng có Hướng dẫn bởi Khả năng Giải thích. Luồng huấn luyện bao gồm: (1) Refiner tinh chỉnh ảnh giả mạo. (2) Detector (Face Encoder và các nhánh đầu ra) phân loại ảnh thật/giả và dự đoán các thành phần khuôn mặt. (3) Prediction Explainer tạo bản đồ CAM từ kết quả phân loại. (4) Bản đồ CAM được sử dụng làm tín hiệu phản hồi để hướng dẫn Refiner trong vòng lặp tiếp theo.	28
4.1	Hình ảnh minh họa về tập dữ liệu FakeAVCeleb.	34

DANH SÁCH BẢNG

- | | | |
|-----|--|----|
| 4.1 | So sánh hiệu suất của các nghiên cứu trên tập dữ liệu FakeAVCeleb. | 36 |
| 4.2 | Kết quả nghiên cứu tách biệt, thể hiện sự ảnh hưởng của các thành phần trong mô hình trên tập dữ liệu FakeAVCeleb. . . . | 37 |

TÓM TẮT ĐỀ TÀI

Sự phát triển nhanh chóng của công nghệ Deepfake đặt ra những thách thức nghiêm trọng đối với an ninh thông tin và mức độ tin cậy của nội dung số. Mặc dù các mô hình phát hiện hiện tại đạt hiệu quả cao trên các tập dữ liệu quen thuộc, chúng thường gặp khó khăn trong việc tổng quát hóa đối với các kỹ thuật giả mạo mới và chưa từng thấy.

Trong nghiên cứu này, chúng tôi đề xuất một phương pháp mới nhằm tăng cường độ bền vững (robustness) và khả năng tổng quát hóa của các mô hình phát hiện Deepfake, thông qua một cơ chế có tên gọi "trò chơi đối kháng có hướng dẫn giải thích" (explainability-guided adversarial game). Phương pháp được triển khai trong một kiến trúc bao gồm hai thành phần chính: Mô hình Phát hiện (Detector) và Mô hình Tinh chỉnh Giả mạo (Refiner).

Điểm nổi bật của phương pháp là vòng lặp phản hồi giữa hai mô hình. Cụ thể, mô hình Detector không chỉ thực hiện nhiệm vụ phân loại hình ảnh thật/giả mà còn học đồng thời các đặc trưng nhận diện khuôn mặt. Dựa trên kết quả phân loại, một bản đồ kích hoạt lớp (Class Activation Map - CAM) được sinh ra nhằm chỉ ra các vùng hình ảnh đóng vai trò quyết định trong việc nhận diện giả mạo. Bản đồ CAM này sau đó được cung cấp cho mô hình Refiner, để thực hiện tinh chỉnh hình ảnh - cụ thể là "xóa bỏ" hoặc che giấu các dấu vết giả mạo tại chính những vùng đã được chỉ định.

Cơ chế cạnh tranh giữa hai mô hình tạo nên một tiến trình huấn luyện đối kháng, trong đó mô hình Detector buộc phải học cách phát hiện các đặc trưng tinh vi, bền vững hơn, thay vì phụ thuộc vào các tín hiệu giả mạo dễ bị loại bỏ. Kết quả thực nghiệm cho thấy phương pháp đề xuất đạt kết quả tiềm năng so với các kỹ thuật hiện có. Điều này cho thấy tiềm năng rõ rệt trong việc cải thiện tính bền vững và năng lực tổng quát hóa của hệ thống phát hiện Deepfake.

CHƯƠNG 1

TỔNG QUAN

1.1 Giới thiệu vấn đề

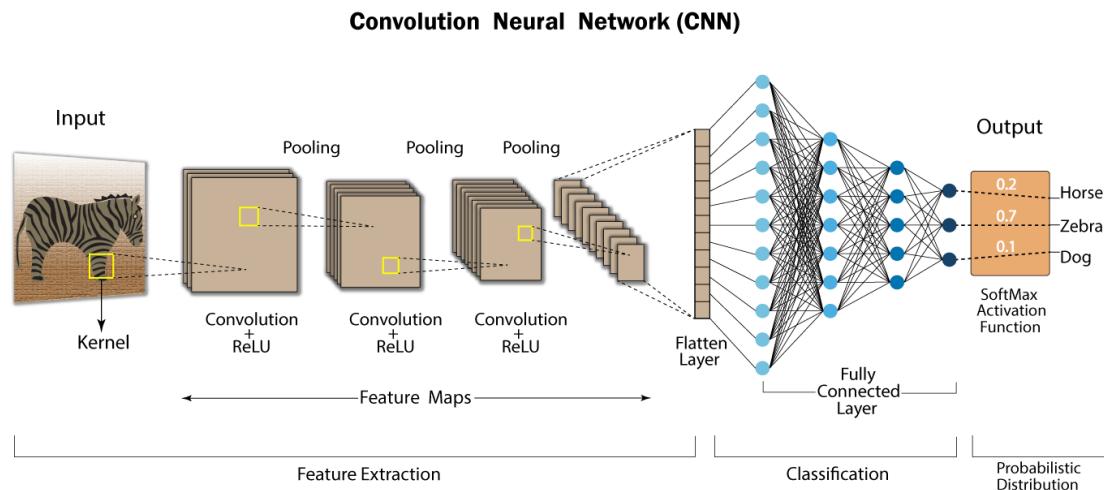
Sự phát triển mạnh của các mô hình học sâu, điển hình là Mạng đối kháng tạo sinh (Generative Adversarial Networks - GAN) [1] và gần đây là các Mô hình Khuếch tán (Diffusion Models) [2], đã tạo ra một cuộc cách mạng trong lĩnh vực Trí tuệ Nhân tạo Tạo sinh (Generative AI) [3]. Các mô hình này sở hữu khả năng phi thường trong việc tổng hợp dữ liệu đa phương tiện bao gồm hình ảnh, video, và âm thanh,... với mức độ chân thực đến mức khó có thể phân biệt bằng mắt thường. Tiềm năng ứng dụng của chúng vô cùng rộng lớn, trải dài từ các lĩnh vực sáng tạo như nghệ thuật số, sản xuất phim, đến các ngành khoa học như mô phỏng y tế và tăng cường dữ liệu.



Hình 1.1: Ảnh minh họa việc các đối tượng lừa đảo sử dụng công nghệ Deepfake nhằm thực hiện các hành vi mạo danh người khác cho các mục đích xấu.

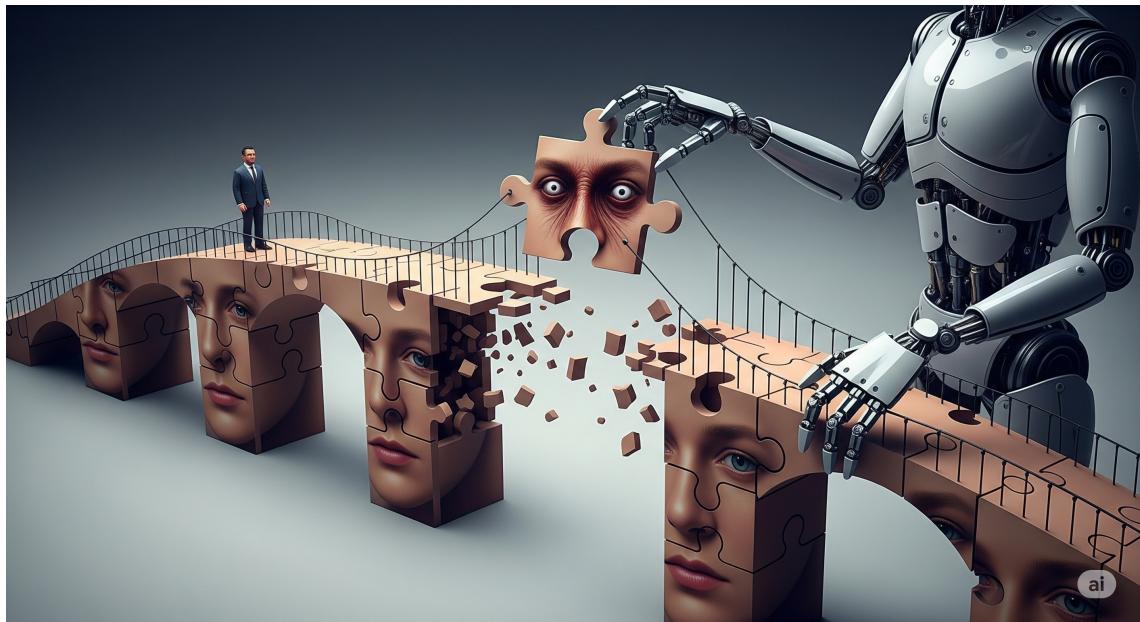
Tuy nhiên, song hành với những tiến bộ này là sự trỗi dậy của một mối đe dọa nghiêm trọng đó là công nghệ Deepfake [4]. Bằng cách lạm dụng sức mạnh của AI tạo sinh, những kẻ xấu có thể tạo ra các sản phẩm truyền thông tổng hợp (synthetic media) để hoán đổi hoặc tái tạo khuôn mặt và giọng nói

một cách vô cùng tinh vi. Sự mở rộng của Deepfake đã và đang gây ra những hậu quả sâu sắc, làm suy giảm sự tin cậy trong xã hội số. Các mối nguy hiểm chính mà deepfake gây ra như: (i) Thao túng thông tin và can thiệp chính trị, nơi các video giả mạo của các nhà lãnh đạo có thể được sử dụng để kích động bất ổn xã hội hoặc gây ảnh hưởng đến các cuộc bầu cử [5]; (ii) Lừa đảo tài chính và mạo danh, khi các hệ thống xác thực sinh trắc học dựa trên khuôn mặt và giọng nói có thể bị qua mặt [6]; và (iii) Bôi nhọ danh dự và tấn công cá nhân, gây ra những tổn thương tâm lý không thể khắc phục cho các nạn nhân [7]. Hơn nữa, sự tồn tại của Deepfake có nguy cơ tạo ra một "nghịch lý của kẻ lừa dối" (liar's dividend), nơi ngay cả những bằng chứng hình ảnh và video chân thực cũng có thể bị bác bỏ vì bị nghi ngờ là giả mạo, làm suy yếu nghiêm trọng vai trò của truyền thông và hệ thống pháp luật.



Hình 1.2: Sơ đồ minh họa kiến trúc tổng quát của một mô hình CNN

Để đối phó với thách thức này, việc phát triển các công cụ phát hiện Deepfake tự động trở nên vô cùng quan trọng giúp bảo vệ sự thật, duy trì an ninh và bảo vệ các cá nhân khỏi các cuộc tấn công dựa trên danh tính. Các phương pháp dựa trên Mạng Nơ-ron Tích chập (CNNs) đã trở thành hướng tiếp cận chủ đạo, cho thấy hiệu quả đáng kể trong việc nhận diện các "dấu vết giả mạo" (forgery artifacts) tinh vi do quá trình tổng hợp để lại. Các dấu vết này có thể tồn tại ở nhiều dạng khác nhau, từ sự thiếu nhất quán trong miền không gian (spatial inconsistencies) [8], các mẫu bất thường trong miền tần số (frequency-domain artifacts) [9], cho đến sự thiếu tự nhiên trong các tín hiệu sinh học (biological signals) [10].



Hình 1.3: Hình ảnh minh họa một robot đang cố gắng xây một cây cầu từ những mảnh ghép chỉ có hình dạng của một loại khuôn mặt duy nhất. Cây cầu bắt đầu sụp đổ ở điểm mà một mảnh ghép có hình dạng khuôn mặt mới, khác biệt không thể khớp vào. Điều này tượng trưng cho việc AI đã "quá khớp" vào dữ liệu huấn luyện và thất bại trong việc tổng quát hóa khi gặp phải một loại dữ liệu mới mà nó chưa từng được học.

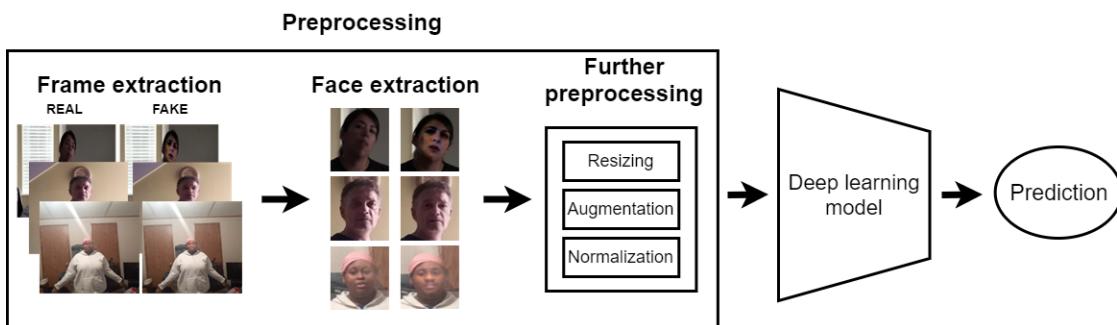
Mặc dù đã đạt được những thành công ban đầu, các mô hình phát hiện hiện tại phải đối mặt với một thách thức vô cùng nghiêm trọng: sự thiếu bền vững (lack of robustness) và khả năng tổng quát hóa (poor generalization). Các mô hình này thường có xu hướng "quá khớp" (overfit) trên các dấu vết đặc trưng của các phương pháp tạo giả có trong tập dữ liệu huấn luyện. Hệ quả là, hiệu suất của chúng suy giảm đáng kể khi phải đối mặt với: (a) các kỹ thuật Deepfake chưa từng thấy (unseen manipulation techniques), ví dụ như một mô hình được huấn luyện trên dữ liệu GAN sẽ hoạt động kém trên dữ liệu từ Diffusion Model; (b) các phép biến đổi hậu kỳ phổ biến (common post-processing operations) như nén ảnh/video, thay đổi kích thước, hoặc thêm nhiễu, vốn có thể làm suy yếu hoặc xóa bỏ các dấu vết mà mô hình dựa vào; và (c) các cuộc tấn công đối kháng (adversarial attacks) được thiết kế có chủ đích để đánh lừa bộ phân loại [11]. Nguyên nhân sâu xa của sự mong manh này là do các mô hình thường học các "lỗi tắt" — chúng dựa vào các tương quan giả (spurious correlations) và các dấu vết cấp thấp, thay vì học các đặc trưng ngữ nghĩa về cấu trúc khuôn mặt thật và giả.

Để giải quyết những hạn chế trên, nghiên cứu này đề xuất một hướng tiếp cận mới nhằm nâng cao tính bền vững và khả năng tổng quát hóa của các mô

hình phát hiện Deepfake. Thay vì huấn luyện một bộ phát hiện (Detector) một cách thụ động, chúng tôi đề xuất một khung Trò chơi Đối kháng có Hướng dẫn bởi Khả năng Giải thích (Explainability-Guided Adversarial Game). Trong khung làm việc này, Detector phải cạnh tranh với một tác nhân đối kháng, được gọi là mô hình Tinh chỉnh (Refiner). Nhiệm vụ của Refiner là nhận một ảnh Deepfake và tinh chỉnh nó để xóa bỏ các dấu vết có thể bị phát hiện. Điểm cốt lõi và mới lạ trong phương pháp của chúng tôi là quá trình tinh chỉnh này không diễn ra một cách "mù quáng". Thay vào đó, nó được dẫn dắt trực tiếp bởi phản hồi từ chính Detector dưới dạng bản đồ giải thích (explainability map), cụ thể là Bản đồ Kích hoạt Lớp (Class Activation Maps - CAM) [12]. Bằng cách buộc Refiner tập trung tấn công vào các vùng mà Detector cho là "đáng ngờ" nhất, chúng tôi tạo ra một cuộc cạnh tranh có mục tiêu. Vòng lặp phản hồi này buộc Detector phải từ bỏ việc dựa dẫm vào các dấu vết bì mặt, dễ bị khai thác, và thay vào đó học các đặc trưng sâu sắc hơn, bền vững hơn và có ý nghĩa ngữ nghĩa hơn.

1.2 Các công trình nghiên cứu liên quan

Lĩnh vực pháp y hình ảnh Deepfake đã phát triển nhanh chóng, với các phương pháp tiếp cận có thể được phân loại thành ba hướng chính: (i) các phương pháp dựa trên phân tích dấu vết giả mạo (artifact-based methods), (ii) các phương pháp khai thác tín hiệu sinh học và vật lý (biological/physical signal-based methods), và (iii) các phương pháp tập trung vào việc tăng cường tính bền vững và khả năng tổng quát hóa (robustness and generalization enhancement methods).



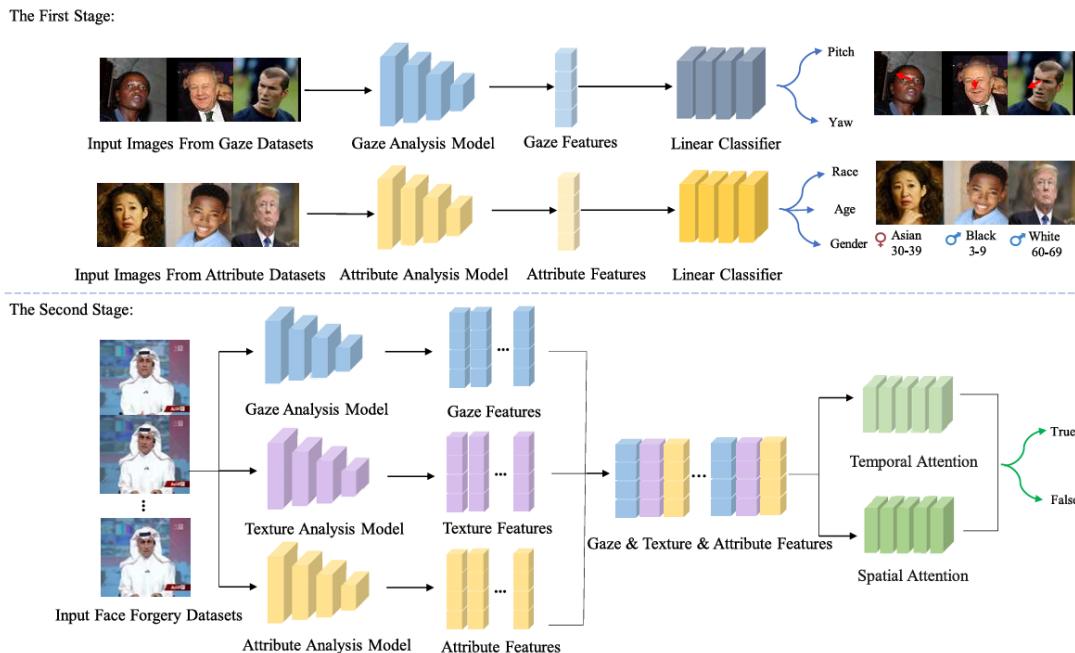
Hình 1.4: Luồng hoạt động cơ bản của các mô hình deepfake detection. Các hình ảnh có thể được phân tích đặc trưng tần số trước hoặc được đưa trực tiếp vào mô hình DNN. Các này phụ thuộc quá mức vào các dấu vết đặc thù của một dataset nên có xu hướng bị overfitting.

Phát hiện Dựa trên Dấu vết Giả mạo trong Miền Không gian

và Tân số. Các phương pháp ban đầu và phổ biến nhất tập trung vào việc xác định các dấu vết giả mạo (forgery artifacts) do quá trình tổng hợp của mô hình sinh để lại. Trong miền không gian, các mô hình như XceptionNet [4] và MesoNet [8] đã được huấn luyện để nhận diện các điểm bất thường ở cấp độ pixel, chẳng hạn như ranh giới ghép nối không nhất quán, hiện vật do nén, hoặc sự thiếu tự nhiên về ánh sáng và đổ bóng. Trong miền tần số, các nghiên cứu đã chỉ ra rằng quá trình nâng mẫu (up-sampling) trong kiến trúc GAN thường để lại các "vết seo" (artifacts) có thể nhận biết trong phô tần số [9]. Các phương pháp này phân tích hình ảnh sau khi qua biến đổi Fourier để phát hiện các mẫu tần số cao bất thường, vốn là một dấu hiệu đáng tin cậy của hình ảnh được tạo ra bằng máy tính [13]. Hạn chế cốt lõi của các phương pháp này là sự phụ thuộc quá mức vào các dấu vết đặc thù của một kỹ thuật tạo giả cụ thể. Chúng có xu hướng "quá khớp" (overfit) trên các hiện vật này, dẫn đến hiệu suất suy giảm nghiêm trọng khi đối mặt với các phương pháp Deepfake mới hoặc khi các hình ảnh đã qua xử lý hậu kỳ (ví dụ: nén, thay đổi kích thước), làm mờ hoặc xóa bỏ các dấu vết đó [4].

Phát hiện Dựa trên Tín hiệu Sinh học và Vật lý. Để vượt qua sự phụ thuộc vào các dấu vết cấp thấp, một hướng nghiên cứu khác tập trung vào việc khai thác các tín hiệu sinh học và vật lý mà các mô hình tạo sinh khó có thể tái tạo một cách hoàn hảo. Các phương pháp này phân tích các hành vi vô thức của con người như tần suất chớp mắt không đều [10], nhịp tim phản ánh qua sự thay đổi màu da vi tế (photoplethysmography) [14], hoặc các chuyển động đầu và biểu cảm khuôn mặt không tự nhiên. Gần đây, Peng và cộng sự (2024) trong bài báo "Where Deepfakes Gaze at?" [15] đã đề xuất một hướng đi độc đáo bằng cách phân tích sự thiếu nhất quán trong hướng nhìn (gaze inconsistency) theo không gian-thời gian. Họ nhận thấy rằng trong các video thật, hướng nhìn của một người có xu hướng tập trung và nhất quán trong một khoảng thời gian ngắn, trong khi các video Deepfake thường thể hiện sự phân tán và thiếu ổn định trong hướng nhìn. Bằng cách kết hợp phân tích hướng nhìn với các đặc trưng về thuộc tính (tuổi, giới tính) và kết cấu da, họ đã đạt được kết quả ấn tượng. Mặc dù rất hiệu quả trong việc nắm bắt các tín hiệu tinh vi, các phương pháp này vẫn có những điểm yếu. Thứ nhất, chúng dễ bị tổn thương trước sự cải tiến không ngừng của các mô hình tạo sinh; trong tương lai, các mô hình này hoàn toàn có thể học cách tạo ra các tín hiệu sinh học nhất quán hơn. Thứ hai, và quan trọng hơn, các phương pháp này được thiết kế để phát hiện các dấu vết "tự nhiên" của quá trình giả mạo, chứ không được thiết kế để chống lại các cuộc tấn công đối kháng có

chủ đích (deliberate adversarial attacks). Một kẻ tấn công có thể tạo ra các nhiễu loạn nhỏ để đánh lừa bộ phân loại mà không cần phải tạo ra hướng nhìn hoàn hảo.



Hình 1.5: "Where Deepfakes Gaze at?". Nghiên cứu này đã tận dụng những đặc trưng về các thành phần trên gương mặt giúp mô hình khai thác được nhiều khía cạnh hơn.

Các phương pháp Tăng cường Tính Bền vững và Khả năng Tổng quát hóa. Nhận thức được những hạn chế của các phương pháp trên, các nghiên cứu gần đây đã chuyển sang việc xây dựng các mô hình có tính bền vững và khả năng tổng quát hóa cao hơn. Hướng đi chủ đạo trong nhóm này là huấn luyện đối kháng (adversarial training). Một công trình tiêu biểu trong hướng này là của Wang và cộng sự (2022) trong bài báo "Deepfake Forensics via an Adversarial Game" [16]. Họ đề xuất một cơ chế huấn luyện đối kháng, nơi một bộ phát hiện (detector) được huấn luyện để chống lại các mẫu giả mạo được tạo ra để tấn công chính nó. Điểm mới lạ trong công trình của họ là việc đề xuất một phương pháp tấn công đối kháng dựa trên việc làm mờ (adversarial blurring), buộc mô hình phải học các đặc trưng ít phụ thuộc vào các hiện vật tần số cao. Họ cũng sử dụng một mạng sinh (generator) để tạo ra các mẫu tấn công này một cách hiệu quả. Mặc dù phương pháp của Wang và cộng sự là một bước tiến quan trọng trong việc tăng cường tính bền vững, cuộc "chạy đua vũ trang" của họ vẫn còn những hạn chế. Cuộc tấn công đối kháng chủ yếu dựa trên gradient của hàm mất mát (gradient-based), có thể được xem là một cuộc tấn công "mù quáng" (blind attack). Nó tối đa hóa

sai số của bộ phát hiện nhưng không cung cấp thông tin chi tiết về lý do tại sao bộ phát hiện bị đánh lừa ở cấp độ ngữ nghĩa. Do đó, bộ phát hiện có thể chỉ học cách chống lại các nhiễu loạn thông kê thay vì thực sự hiểu các đặc trưng cấu trúc cơ bản của một khuôn mặt. Hơn nữa, phương pháp này không tường minh buộc mô hình phải học các đặc trưng ngữ nghĩa cao (high-level semantic features) như vị trí của mắt, mũi, miệng.

1.3 Mục tiêu, đối tượng, phạm vi nghiên cứu

1.3.1 Mục tiêu nghiên cứu

Nghiên cứu này đề xuất và triển khai một phương pháp mới nhằm nâng cao tính bền vững (robustness) và khả năng tổng quát hóa (generalization) cho các mô hình phát hiện hình ảnh Deepfake. Các mục tiêu cụ thể bao gồm:

- **Xây dựng khung huấn luyện đối kháng có hướng dẫn:** Đề xuất và hiện thực hóa kiến trúc "Trò chơi Đối kháng có Hướng dẫn bởi Khả năng Giải thích" (Explainability-Guided Adversarial Game - EAG). Trong đó, một Mô hình Tinh chỉnh (Refiner) học cách tạo ra các mẫu giả mạo tinh vi hơn bằng cách tận dụng tín hiệu phản hồi từ Mô hình Phát hiện (Detector).
- **Tăng cường học đặc trưng ngữ nghĩa:** Tích hợp một nhánh học đa nhiệm (multi-task learning) vào Detector, với nhiệm vụ phụ là phát hiện các thành phần trên khuôn mặt (mắt, mũi, miệng). Mục tiêu là buộc mô hình phải học các đặc trưng có ý nghĩa về cấu trúc khuôn mặt, thay vì chỉ dựa vào các dấu vết giả mạo cấp thấp.
- **Thực nghiệm và đánh giá hiệu quả:** Triển khai mô hình, tiến hành huấn luyện và đánh giá hiệu suất trên một tập dữ liệu chuẩn (benchmark dataset). So sánh kết quả với các phương pháp hiện đại và thực hiện các nghiên cứu tách biệt (ablation studies) để chứng minh sự đóng góp của từng thành phần trong phương pháp đề xuất.

1.3.2 Đối tượng nghiên cứu

Vấn đề: Bài toán phát hiện và phân loại hình ảnh khuôn mặt giả mạo (Deepfake image detection).

Phương pháp: Các kỹ thuật học sâu liên quan, bao gồm:

- Huấn luyện đối kháng (Adversarial Training).
- Học đa nhiệm (Multi-task Learning).
- Các phương pháp giải thích mô hình (Explainable AI - XAI), cụ thể là Class Activation Map (CAM).

Dữ liệu: Các hình ảnh khuôn mặt thật và giả mạo được trích xuất từ các tập dữ liệu video công khai.

1.3.3 Phạm vi nghiên cứu

Về dữ liệu: Nghiên cứu tập trung vào việc xử lý dữ liệu hình ảnh (các khung hình được trích xuất từ video), không đi sâu vào việc phân tích đồng thời luồng âm thanh (audio) hay các tín hiệu đa phương thức khác. Tập dữ liệu chính được sử dụng cho thực nghiệm là FakeAVCeleb.

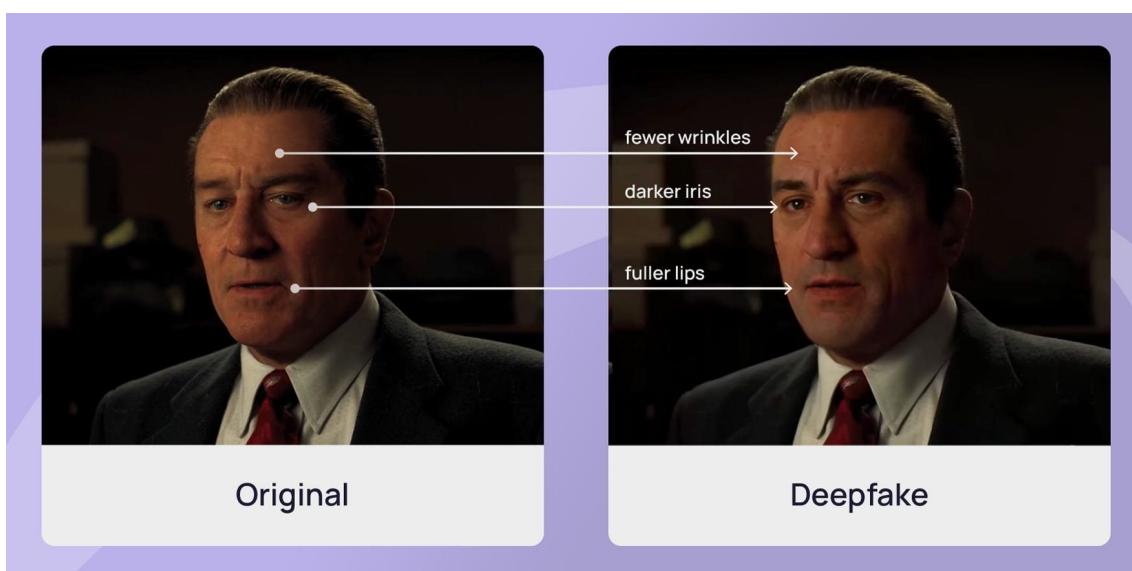
Về phương pháp: Đề tài tập trung vào việc phát triển và đánh giá khung huấn luyện EAG được đề xuất. Nghiên cứu sử dụng CAM làm kỹ thuật giải thích chính và không đi sâu vào việc so sánh toàn diện tất cả các kỹ thuật XAI khác nhau (như Grad-CAM, LIME, SHAP).

Về ứng dụng: Đồ án tập trung vào khía cạnh kỹ thuật của bài toán phát hiện, nhằm mục đích xây dựng một mô hình có hiệu suất cao và bền vững. Các khía cạnh về pháp lý, đạo đức, hay triển khai hệ thống trong thực tế không thuộc phạm vi nghiên cứu chính của đồ án này.

CHƯƠNG 2

CƠ SỞ LÝ THUYẾT

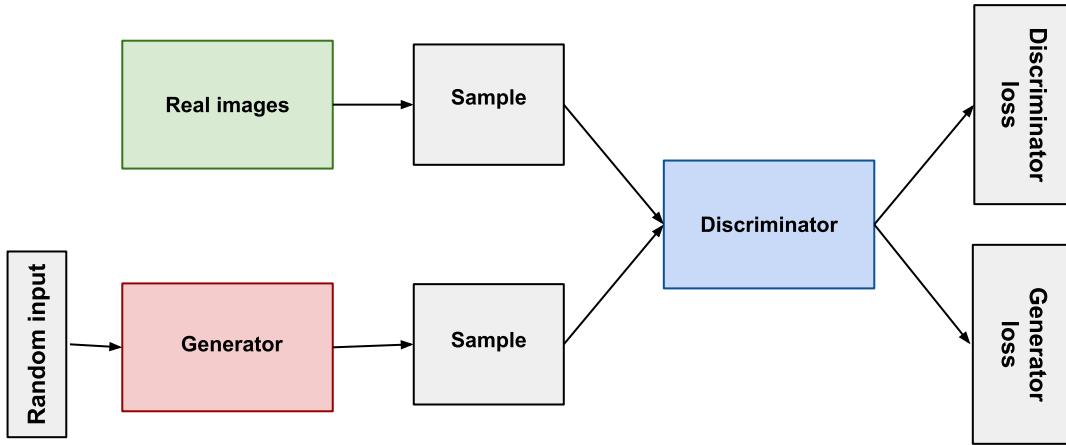
Chương này trình bày các cơ sở lý thuyết nền tảng, làm tiền đề cho việc nghiên cứu và phát triển phương pháp được đề xuất trong đồ án. Để hiểu rõ bản chất hình ảnh Deepfake và các thách thức trong việc phát hiện, trước tiên cần nắm vững kiến trúc và nguyên lý hoạt động của Mạng đối nghịch tạo sinh (GANs) – công nghệ cốt lõi đứng sau các kỹ thuật tạo Deepfake tinh vi hiện nay. Sự phát triển nhanh chóng của các công nghệ tạo giả đã thúc đẩy sự ra đời của các phương pháp phát hiện Deepfake, nhằm phân biệt nội dung thật và giả mạo.



Hình 2.1: So sánh ảnh gốc và ảnh Deepfake. Các đặc điểm khuôn mặt đã bị chỉnh sửa tinh vi, gây khó khăn cho việc phát hiện.

Do bản chất đối kháng giữa quá trình tạo và phát hiện, lý thuyết huấn luyện đối kháng đóng vai trò quan trọng, được xem là một trong những cơ chế phòng thủ hiệu quả giúp nâng cao độ tin cậy cho các mô hình phát hiện. Bên cạnh đó, để vượt qua tính chất "hộp đen" của các mô hình học sâu, các phương pháp dựa trên khả năng giải thích cũng sẽ được phân tích. Những phương pháp này không chỉ hỗ trợ đánh giá mô hình mà còn giúp định hướng cải tiến, phù hợp với mục tiêu chính của đồ án. Các nội dung trên tạo thành một chuỗi logic – từ công nghệ nền tảng, ứng dụng đến giải pháp phòng thủ – làm nền tảng vững chắc cho các chương tiếp theo.

2.1 GANs (Generative Adversarial Networks)



Hình 2.2: Tổng quan về kiến trúc của mạng GAN (Generative Adversarial Network). Kiến trúc chính bao gồm một Generator và một Discriminator cạnh tranh nhau trong quá trình huấn luyện.

Mạng đối nghịch tạo sinh (GANs)[1] là một lớp các mô hình học máy gồm hai mạng nơ-ron cạnh tranh với nhau. Cấu trúc này được thiết kế để tạo ra các mẫu dữ liệu mới có đặc điểm thống kê tương tự như tập dữ liệu huấn luyện. Hai thành phần chính của GAN là:

- **Mạng tạo sinh (Generator - G):** Mạng này học cách tạo ra dữ liệu giả mạo. Nó nhận một véc-tơ nhiễu ngẫu nhiên làm đầu vào và cố gắng biến đổi nó thành một mẫu dữ liệu (ví dụ: một bức ảnh khuôn mặt) trông giống thật nhất có thể.
- **Mạng phân biệt (Discriminator - D):** Mạng này hoạt động như một bộ phân loại, có nhiệm vụ xác định xem dữ liệu đầu vào là "thật" (lấy từ tập dữ liệu thực tế) hay "giả" (do Generator tạo ra).

Trong quá trình huấn luyện, hai mạng này tham gia vào một trò chơi đối kháng. Generator liên tục cải thiện để tạo ra những sản phẩm giả mạo ngày càng thuyết phục nhằm đánh lừa Discriminator. Ngược lại, Discriminator cũng liên tục được huấn luyện để trở nên nhạy bén hơn trong việc phát hiện ra sản phẩm giả. Quá trình cạnh tranh này thúc đẩy cả hai mạng cùng phát triển, cho đến khi Generator có thể tạo ra những sản phẩm giả mạo chất lượng cao, rất khó để phân biệt.

2.2 Deepfake Generation

Deepfake Generation là quá trình sử dụng các mô hình học sâu để tạo ra hình ảnh hoặc video giả mạo, thường dựa trên GANs và các biến thể của nó. Các kỹ thuật chính được chia thành ba nhóm: Entire Face Synthesis, Face Identity Swap, và Face Manipulation.



Hình 2.3: Hình ảnh các khuôn mặt được tạo sinh hoàn toàn bằng mô hình StyleGAN của NVIDIA. Dù trông giống người thật, tất cả đều là khuôn mặt giả (synthetic faces), không tồn tại trong thực tế.

2.2.1 Entire Face Synthesis (Tổng hợp toàn bộ khuôn mặt)

Nhóm kỹ thuật này tập trung vào việc tạo ra các khuôn mặt hoàn toàn mới, không tồn tại trong thực tế, với độ chân thực cao.

- **ProGAN (Progressive Growing of GANs)[17]:** Phương pháp này tăng dần độ phân giải của hình ảnh được tạo ra, bắt đầu từ độ phân giải thấp và bổ sung chi tiết dần dần. Điều này cải thiện chất lượng và độ ổn định của hình ảnh giả.
- **StyleGAN[18]:** Được phát triển bởi NVIDIA, StyleGAN sử dụng kiến trúc dựa trên phong cách (style-based), cho phép kiểm soát tốt hơn các thuộc tính của hình ảnh như đặc điểm khuôn mặt, biểu cảm, và hậu

cảnh. Ví dụ các hình ảnh được tạo ra bởi StyleGAN được thể hiện trong Hình 2.3.

2.2.2 Face Identity Swap (Hoán đổi danh tính khuôn mặt)

Kỹ thuật này thay thế khuôn mặt của một người bằng khuôn mặt của người khác, giữ nguyên biểu cảm và chuyển động ban đầu.

- **FaceSwap:** Một công cụ phổ biến sử dụng GANs để thực hiện hoán đổi danh tính, duy trì tính tự nhiên của video hoặc hình ảnh.
- **FSGAN (Face Swapping GAN)[19]:** Tập trung vào hoán đổi danh tính không phụ thuộc vào đối tượng cụ thể, mang lại kết quả linh hoạt và chân thực hơn.



Hình 2.4: Minh họa hoán đổi danh tính khuôn mặt sử dụng FSGAN. Khuôn mặt từ ảnh gốc (source image) được hoán đổi lên video mục tiêu (target video), kết quả nằm ở giữa.

2.2.3 Face Manipulation (Thao tác khuôn mặt)

Nhóm này chỉnh sửa các thuộc tính cụ thể của khuôn mặt như màu tóc, giới tính, hoặc tuổi tác.

- **StarGAN[20]:** Một mô hình thông nhất cho việc chuyển đổi hình ảnh đa miền, có khả năng thay đổi nhiều thuộc tính cùng lúc.
- **AttGAN[21]:** Tập trung vào chỉnh sửa thuộc tính bằng cách thay đổi các đặc điểm cụ thể trong khi giữ nguyên các phần khác.
- **STGAN[22]:** Một mạng chuyển giao chọn lọc cho phép chỉnh sửa thuộc tính tùy ý với mức độ kiểm soát chi tiết.

Các kỹ thuật này tận dụng sức mạnh của GANs để tạo ra hình ảnh hoặc video giả có độ chân thực cao, khó phân biệt bằng mắt thường.

2.3 Deepfake Detection

Deepfake Detection là một bài toán nhị phân nhằm phân biệt đối tượng thật và giả mạo. Các phương pháp phát hiện dựa trên nguyên tắc rằng trình tạo ảnh của AI, dù tinh vi những vẫn để lại những dấu vết giả mạo (artifacts) hoặc những sự khác biệt thống kê so với ảnh tự nhiên. Các máy dò được huấn luyện để nhận diện những dấu vết này.

Các phương pháp phát hiện thường tập trung vào việc phân tích sự khác biệt thống kê giữa hình ảnh thật và giả trong cả miền không gian (spatial domain) và miền tần số (frequency domain).

2.3.1 Spatial Domain Detection (Phát hiện trong miền không gian)

Các phương pháp này phân tích trực tiếp dữ liệu pixel của hình ảnh để tìm kiếm các dấu vết giả mạo. Các dấu vết này có thể là sự thiếu tự nhiên về màu sắc mắt, các chi tiết bị mờ ở răng, hoặc các điểm không nhất quán về kết cấu trên da.

Một số phương pháp tiếp cận:

- **Phương pháp dựa trên CNN:** Sử dụng mạng nơ-ron tích chập (Convolutional Neural Networks - CNN) như ResNet [23], EfficientNet [24], DenseNet [25], và MobileNet [26] để trích xuất đặc trưng từ hình ảnh và phân loại chúng là thật hay giả.
- **Phương pháp dựa trên phân tích từng mảng (Patch-based) [27]:** Tập trung phân tích các mảng nhỏ của hình ảnh để phát hiện sự không nhất quán hoặc dấu vết do Deepfake để lại.

2.3.2 Frequency Domain Detection (Phát hiện trong miền tần số)

Các phương pháp này chuyển đổi hình ảnh sang miền tần số để phân tích, vì quá trình tạo ảnh của GAN thường để lại các dấu hiệu đặc trưng trong phổ tần số. Ví dụ, các ảnh do GAN tạo ra có thể có các thành phần tần số cao bất thường hoặc các mẫu lặp lại do quá trình "up-convolution".

Một số phương pháp tiếp cận:

- **Phân tích tần số:** Chuyển đổi hình ảnh sang miền tần số (ví dụ: sử dụng biến đổi Fourier) để phát hiện các bất thường trong thành phần tần số. Các phương pháp như DCTA [28] và DFTD [29] phân tích sự khác biệt tần số để xác định hình ảnh giả.
- **Phân tích phổ (Spectrogram Analysis) [9]:** Kiểm tra phổ của hình ảnh để tìm các dấu vết tần số cao đặc trưng thường xuất hiện trong hình ảnh do GAN tạo ra.

2.4 Adversarial Training

Adversarial Training (Huấn luyện đối kháng)[30] là một trong những phương pháp phòng thủ chủ động và hiệu quả nhất để tăng cường độ tin cậy và sức chống chịu (robustness) cho các mô hình học sâu, đặc biệt là các trình phát hiện Deepfake, trước các cuộc tấn công đối kháng. Về bản chất, đây là một quá trình "tiêm vắc-xin" cho mô hình, bằng cách cho nó tiếp xúc với các phiên bản tấn công được tạo ra một cách có chủ đích, từ đó giúp mô hình học cách nhận biết và trở nên "miễn nhiễm" với chúng.

2.4.1 Nguyên lý hoạt động

Các mô hình học sâu tiêu chuẩn thường được huấn luyện trên các tập dữ liệu "sạch". Điều này khiến chúng học được các mẫu đặc trưng của dữ liệu nhưng lại rất nhạy cảm với những nhiễu loạn nhỏ, có chủ đích mà mắt người không nhận ra. Kẻ tấn công có thể khai thác điểm yếu này để tạo ra các ví dụ đối kháng (adversarial examples), khiến mô hình đưa ra dự đoán sai một cách tự tin.

Huấn luyện đối kháng giải quyết vấn đề này bằng cách đưa chính các ví dụ đối kháng vào quá trình huấn luyện. Mục tiêu không chỉ là để mô hình đạt độ chính xác cao trên dữ liệu sạch, mà còn phải duy trì hiệu suất ổn định khi đối mặt với dữ liệu bị tấn công.

2.4.2 Quy trình hoạt động

Quá trình huấn luyện đối kháng thường là một vòng lặp, bao gồm các bước sau:

1. **Tạo ví dụ đối kháng:** Ở mỗi vòng lặp hoặc mỗi lô (batch) dữ liệu huấn luyện, một thuật toán tấn công đối kháng (ví dụ: Fast Gradient Sign Method - FGSM [30], hay Projected Gradient Descent - PGD [31]) được sử dụng.
 - Lấy một mẫu dữ liệu đầu vào (ví dụ: một ảnh Deepfake).
 - Tính toán gradient của hàm mất mát (loss function) của mô hình hiện tại theo dữ liệu đầu vào đó. Gradient này chỉ ra hướng thay đổi các pixel ảnh để làm tăng độ sai lệch của mô hình nhiều nhất.
 - Tạo ra một nhiễu loạn nhỏ theo hướng gradient đó và cộng nó vào ảnh gốc. Nhiễu loạn này được giới hạn trong một ngưỡng cho phép (thường ký hiệu là ϵ) để đảm bảo mắt người không nhận thấy sự thay đổi.
2. **Tăng cường tập dữ liệu:** Các ví dụ đối kháng vừa được tạo ra sẽ được thêm vào tập dữ liệu huấn luyện. Điều quan trọng là các ví dụ này vẫn giữ nguyên nhãn gốc của chúng. Ví dụ, một ảnh Deepfake sau khi bị biến đổi để tấn công máy dò vẫn được gán nhãn là "giả".
3. **Huấn luyện lại mô hình:** Mô hình máy dò sẽ được cập nhật trọng số dựa trên tập dữ liệu đã được tăng cường, bao gồm cả dữ liệu sạch và dữ liệu đối kháng. Bằng cách này, mô hình buộc phải học cách bỏ qua các nhiễu loạn không liên quan và tập trung vào các đặc điểm bản chất hơn để phân loại.

2.4.3 Mục tiêu và Kết quả

Mục tiêu của huấn luyện đối kháng là giúp mô hình học được một **ranh giới quyết định** vững chắc hơn – thay vì một ranh giới mỏng manh, dễ bị nhiễu loạn vượt qua, mô hình phân định rõ ràng hơn giữa các lớp dữ liệu.

Tuy nhiên, kỹ thuật này vẫn có giới hạn:

- **Hiệu quả với tấn công đã biết:** Mô hình huấn luyện bằng PGD thường chống lại các tấn công PGD rất tốt.

- **Dễ tổn thương với tấn công mới:** Mô hình chỉ chống được những kiểu tấn công đã gặp trong huấn luyện. Nếu tấn công mới thay đổi thuộc tính ngữ nghĩa (không phải pixel), mô hình vẫn có thể bị đánh lừa.

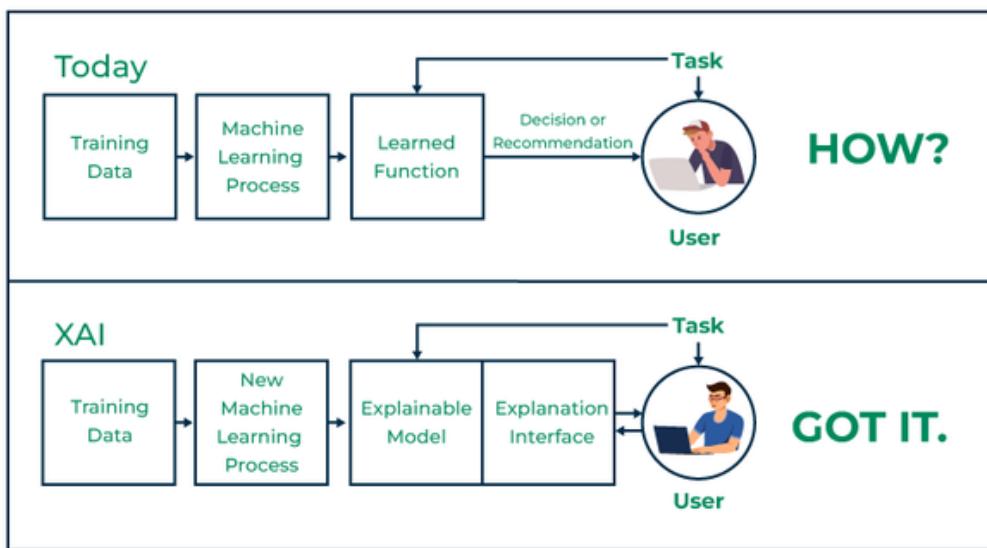
Tóm lại, huấn luyện đối kháng là một kỹ thuật phòng thủ mạnh nhưng không toàn năng. Nó nằm trong cuộc đua không hồi kết giữa tấn công và phòng thủ trong học sâu.

2.5 Explainability-Guided Approach

2.5.1 Khái niệm Explainability trong học sâu

Explainability (khả năng giải thích) trong học sâu (Deep Learning) là khả năng hiểu và diễn giải cách một mô hình đưa ra quyết định. Trong bài toán phát hiện Deepfake, các mô hình học sâu thường hoạt động như một "hộp đen" (black box), nghĩa là rất khó để biết chính xác lý do tại sao mô hình phân loại một hình ảnh là thật hay giả. Điều này gây ra vấn đề:

- Các mô hình truyền thống thường dựa vào các đặc trưng cấp thấp (low-level features) như nhiều pixel hoặc dấu vết nén, nhưng không bền vững khi gặp các kỹ thuật Deepfake mới.
- Việc thiếu khả năng giải thích khiến người dùng không tin tưởng hoàn toàn vào kết quả của mô hình, đặc biệt trong các ứng dụng quan trọng như an ninh hoặc xác thực danh tính.



Hình 2.5: Minh họa tính giải thích (Explainability) trong trí tuệ nhân tạo, nâng cao sự minh bạch và tin cậy của các mô hình học máy

Trong bối cảnh phát hiện Deepfake, **Explainability** đóng vai trò quan trọng vì:

- Giúp hiểu rõ các đặc trưng mà mô hình học được, từ đó đánh giá xem mô hình có tập trung vào các yếu tố phù hợp hay không.
- Tăng độ tin cậy và minh bạch, đặc biệt trong các ứng dụng nhạy cảm như xác thực danh tính.
- Phát hiện lỗi hoặc thiên lệch (bias) trong mô hình, chẳng hạn như việc phụ thuộc quá mức vào các dấu vết giả mạo bề mặt thay vì các đặc trưng ngữ nghĩa sâu sắc.
- Hỗ trợ các chuyên gia kiểm tra và xác thực kết quả, đảm bảo tính chính xác và công bằng.

Ví dụ, trong phát hiện Deepfake, nếu mô hình chỉ dựa vào các yếu tố không liên quan (như nhiễu nền) thay vì đặc trưng khuôn mặt, việc thiếu khả năng giải thích sẽ khiến chúng ta không nhận ra vấn đề này, dẫn đến hiệu suất kém khi gặp dữ liệu mới.

2.5.2 Các phương pháp XAI phổ biến

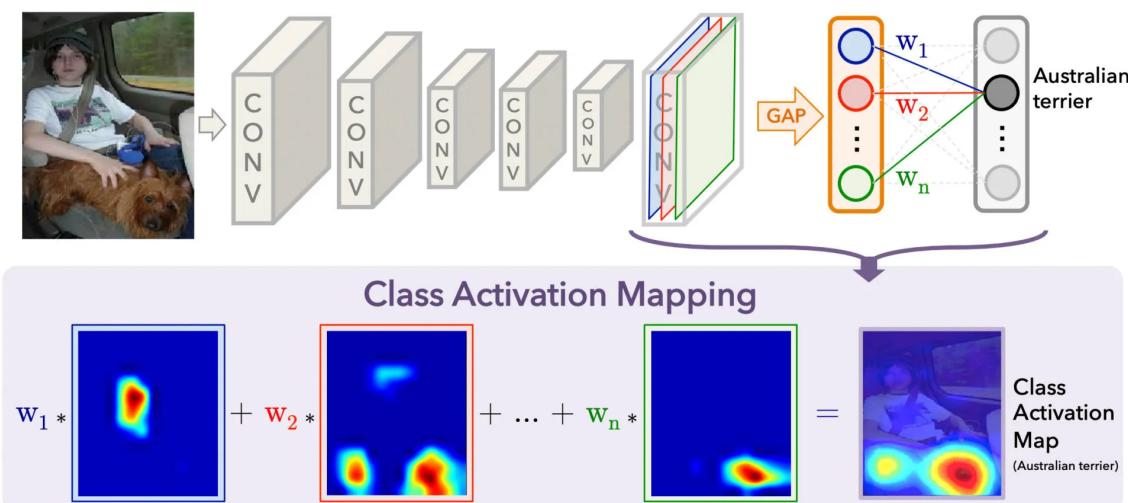
Các mô hình học sâu thường được ví như những "hộp đen" (black boxes) vì quá trình ra quyết định của chúng rất phức tạp và khó diễn giải, đặc biệt trong lĩnh vực xử lý ảnh. Explainable AI (XAI) được ra đời nhằm mục đích làm sáng tỏ lý do đằng sau các dự đoán của mô hình.

Một số phương pháp phổ biến để giải thích mô hình học sâu, đặc biệt trong xử lý ảnh:

- **Saliency Map** [32]: Xác định các vùng ảnh có ảnh hưởng lớn đến quyết định của mô hình.
- **Class Activation Mapping (CAM)** [12]: Tạo bản đồ nhiệt (heatmap) thể hiện các vùng ảnh quan trọng đối với từng lớp dự đoán.
- **Grad-CAM** [33]: Mở rộng CAM bằng cách sử dụng gradient để xác định vùng ảnh quan trọng.
- **LIME** [34], **SHAP** [35]: Các phương pháp giải thích cục bộ, áp dụng cho nhiều loại mô hình.

a. CAM (Class Activation Mapping)

Khái niệm: CAM (Class Activation Mapping) [12] là một kỹ thuật XAI được thiết kế để giải thích các mô hình mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs) trong bài toán phân loại hình ảnh. Phương pháp này tạo ra một bản đồ nhiệt (heatmap) để chỉ ra các vùng quan trọng trong hình ảnh ảnh hưởng đến dự đoán của mô hình cho một lớp cụ thể.



Hình 2.6: Hình ảnh minh họa bản đồ nhiệt (CAM) từ lớp Global Average Pooling trong mạng CNN, thể hiện khả năng giải thích (explainability) của mô hình học sâu trong việc nhận diện và phân tích đặc trưng của hình ảnh, chẳng hạn như một chú chó Australian Terrier trên ghế xe có dây an toàn.

Cách thức hoạt động:

- **Kiến trúc yêu cầu:** CAM yêu cầu mô hình CNN có một lớp Global Average Pooling (GAP) trước lớp đầu ra (fully connected layer). Lớp GAP lấy trung bình giá trị của mỗi bản đồ đặc trưng (feature map) từ lớp tích chập cuối cùng.
- **Tính toán bản đồ nhiệt:**
 1. Lấy các bản đồ đặc trưng từ lớp tích chập cuối cùng.
 2. Nhân các bản đồ đặc trưng này với trọng số của lớp đầu ra tương ứng với lớp dự đoán (class score).
 3. Tổng hợp kết quả để tạo ra bản đồ nhiệt, sau đó chuẩn hóa và chồng lên hình ảnh gốc.
- Kết quả là một bản đồ nhiệt trực quan, trong đó các vùng sáng hơn thể hiện mức độ quan trọng cao hơn đối với dự đoán.

Ví dụ minh họa:

- **Bài toán:** Nhận diện giống chó trong ảnh.
- **Kịch bản:** Mô hình dự đoán một con chó là "Australian Terrier" - Hình 2.6. CAM tạo ra bản đồ nhiệt làm nổi bật các vùng như đầu, tai và đuôi của con chó, cho thấy đây là các đặc trưng chính mà mô hình dựa vào để đưa ra dự đoán.

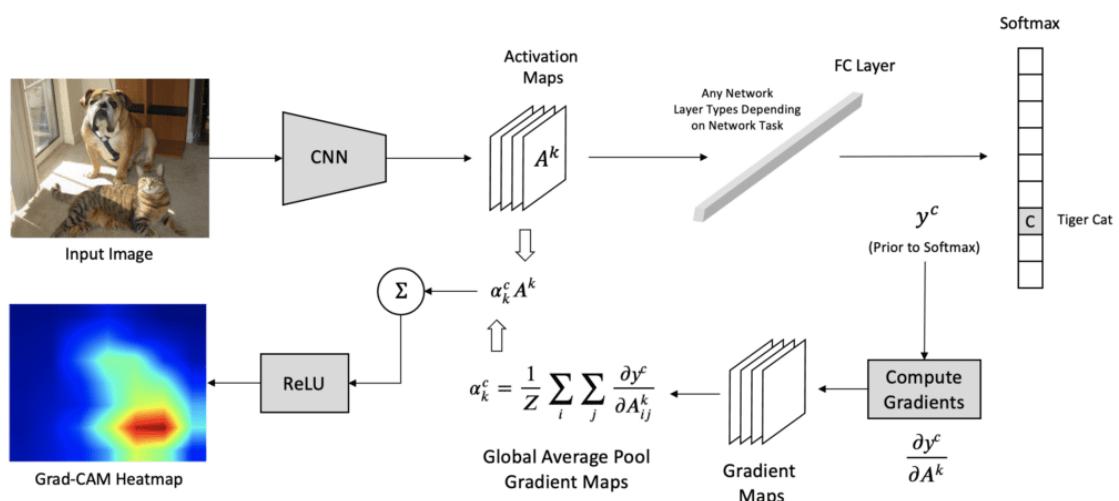
Ưu điểm:

- Dễ triển khai trên các CNN có lớp GAP.
- Trực quan, dễ hiểu với bản đồ nhiệt.

Hạn chế: Chỉ áp dụng được cho các kiến trúc CNN có lớp GAP, không linh hoạt với các mô hình khác. Độ chính xác của bản đồ nhiệt có thể thấp nếu mô hình không được thiết kế phù hợp.

b. Grad-CAM (Gradient-weighted Class Activation Mapping)

Khái niệm: **Grad-CAM**[33] là phiên bản cải tiến của CAM, khắc phục hạn chế về tính linh hoạt của CAM. Phương pháp này sử dụng gradient (đạo hàm) để tạo bản đồ nhiệt, giúp nó áp dụng được cho nhiều kiến trúc CNN khác nhau.



Hình 2.7: Sơ đồ Grad-CAM minh họa cách tạo bản đồ nhiệt từ gradient và lớp kích hoạt CNN, xác định khu vực quan trọng để phân loại hình ảnh (mèo và chó) trước softmax.

Cách thức hoạt động:

- **Bước 1:** Tính gradient của điểm số dự đoán (class score) cho một lớp cụ thể đối với các bản đồ đặc trưng của lớp tích chập cuối cùng.
- **Bước 2:** Lấy trung bình toàn cục (global average) của các gradient này để xác định trọng số tầm quan trọng của từng bản đồ đặc trưng.
- **Bước 3:** Kết hợp các bản đồ đặc trưng với trọng số gradient để tạo bản đồ nhiệt, sau đó áp dụng hàm ReLU để giữ lại các giá trị dương (vùng có ảnh hưởng tích cực đến dự đoán).
- **Kết quả:** Một bản đồ nhiệt chi tiết hơn, làm nổi bật các vùng quan trọng trong hình ảnh.

Ví dụ minh họa

- Bài toán: Phát hiện khối u trong ảnh X-quang.
- Kịch bản: Grad-CAM tạo bản đồ nhiệt chỉ ra chính xác vị trí khối u trên ảnh, giúp bác sĩ xác nhận rằng mô hình đang tập trung vào vùng bất thường thay vì các vùng không liên quan.

Ưu điểm

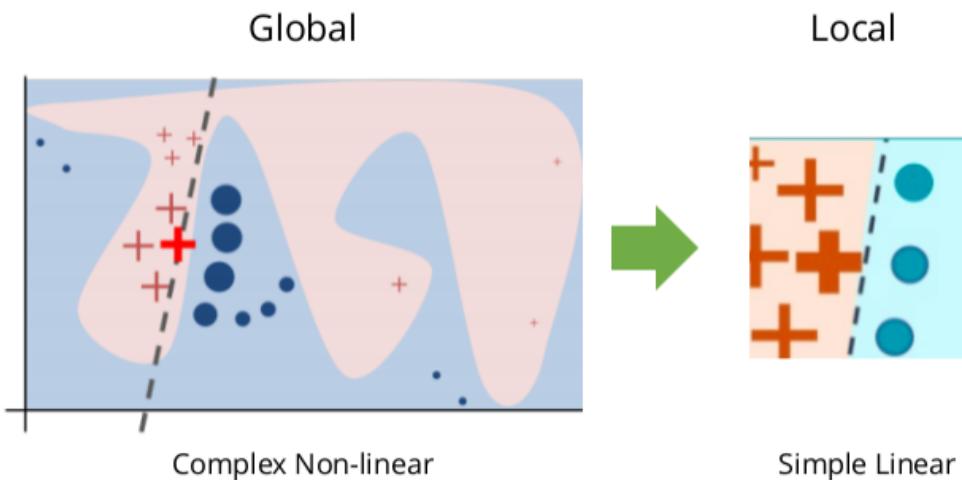
- Linh hoạt, áp dụng được cho hầu hết các kiến trúc CNN.
- Bản đồ nhiệt chính xác hơn nhờ sử dụng gradient.

Hạn chế:

- Yêu cầu tính toán gradient, làm tăng chi phí tính toán so với CAM.
- Kết quả vẫn phụ thuộc vào chất lượng của mô hình gốc.

c. LIME (Local Interpretable Model-agnostic Explanations)

Khái niệm: **LIME** [34] là một phương pháp XAI model-agnostic, nghĩa là nó không phụ thuộc vào loại mô hình học máy cụ thể. LIME tập trung vào việc giải thích các dự đoán cá nhân bằng cách xấp xỉ mô hình phức tạp bằng một mô hình đơn giản hơn, dễ diễn giải.



Hình 2.8: Sơ đồ minh họa phương pháp LIME, chuyển từ mô hình phân loại toàn cục phi tuyến phức tạp sang mô hình cục bộ tuyến tính đơn giản để giải thích.

Cách thức hoạt động:

- **Bước 1:** Tạo các biến thể của đầu vào ban đầu bằng cách thay đổi hoặc loại bỏ một số đặc trưng (ví dụ: che các phần của hình ảnh hoặc thay đổi giá trị trong dữ liệu bảng).
- **Bước 2:** Dự đoán kết quả của các biến thể này bằng mô hình gốc.
- **Bước 3:** Huấn luyện một mô hình đơn giản (như hồi quy tuyến tính hoặc cây quyết định) trên tập dữ liệu biến thể để xấp xỉ hành vi của mô hình gốc xung quanh điểm dữ liệu cần giải thích.
- **Kết quả:** Một giải thích cục bộ, chỉ ra các đặc trưng quan trọng nhất ảnh hưởng đến dự đoán cụ thể.

Ví dụ minh họa

- Bài toán: Phân tích cảm xúc văn bản.
- Kịch bản: Đối với câu "I love this movie, it's amazing!", LIME làm nổi bật các từ "love" và "amazing" là những từ quan trọng dẫn đến dự đoán cảm xúc tích cực.

Ưu điểm:

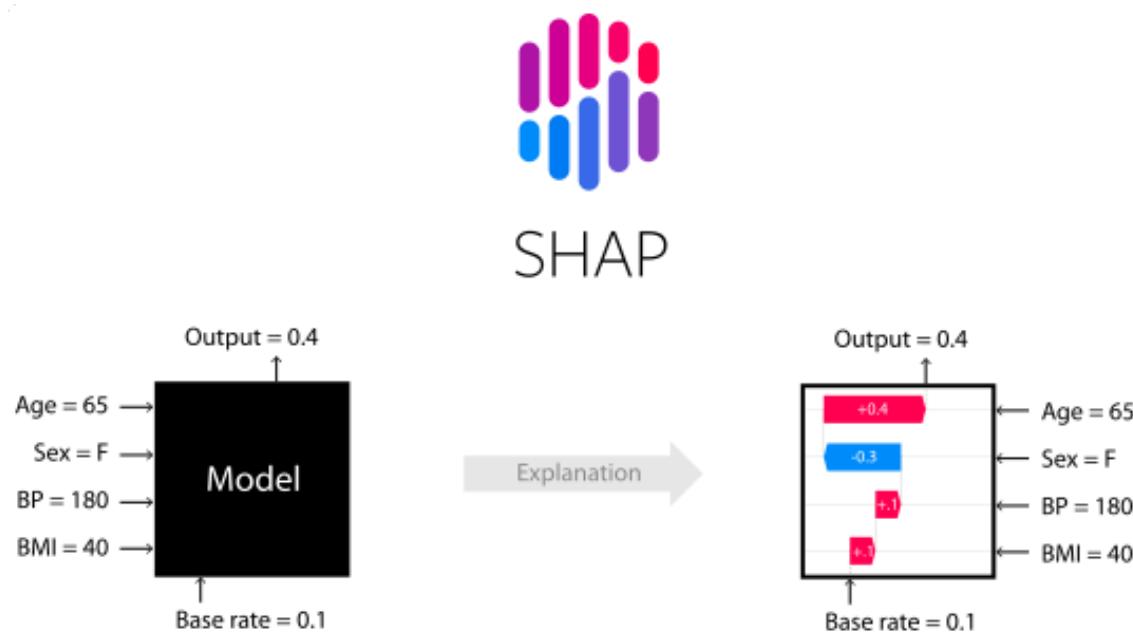
- Linh hoạt, áp dụng được cho bất kỳ loại mô hình nào.
- Cung cấp giải thích cục bộ, dễ hiểu cho từng dự đoán.

Nhược điểm:

- Tốn kém về mặt tính toán do cần tạo và đánh giá nhiều biến thể.
- Giải thích có thể không ổn định nếu các biến thể không đại diện tốt cho dữ liệu gốc.

d. SHAP (SHapley Additive exPlanations)

Khái niệm: SHAP [35] là một phương pháp XAI dựa trên lý thuyết trò chơi, sử dụng giá trị Shapley để gán mức độ quan trọng cho từng đặc trưng trong dự đoán của mô hình. Đây là một trong những phương pháp mạnh mẽ nhất để giải thích cả cục bộ và toàn cục.



Hình 2.9: Sơ đồ minh họa giải thích SHAP, thể hiện tầm quan trọng của các đặc trưng (như tuổi, giới tính, huyết áp, BMI) trong việc đóng góp vào dự đoán và đánh giá mô hình học máy.

Cách thức hoạt động:

- **Nguyên lý:** SHAP dựa trên giá trị Shapley trong lý thuyết trò chơi hợp tác, tính toán mức độ đóng góp trung bình của mỗi đặc trưng bằng cách xem xét tất cả các tổ hợp có thể của các đặc trưng.

- **Bước 1:** Xác định dự đoán cơ sở (baseline prediction) khi không có đặc trưng nào được sử dụng.
- **Bước 2:** Tính toán sự thay đổi trong dự đoán khi thêm từng đặc trưng vào các tổ hợp khác nhau.
- **Bước 3:** Tổng hợp kết quả để gán giá trị SHAP cho từng đặc trưng, cho thấy mức độ ảnh hưởng của nó đến dự đoán cuối cùng.
- **Kết quả:** Một tập hợp giá trị SHAP cho mỗi đặc trưng, có thể được trực quan hóa qua biểu đồ hoặc bảng.

Ví dụ minh họa:

- Bài toán: Dự đoán giá nhà.
- Kịch bản: SHAP cho thấy "vị trí" đóng góp +50.000 USD, "diện tích" đóng góp +30.000 USD, trong khi "màu sắc" chỉ đóng góp +500 USD vào giá dự đoán, giúp người dùng hiểu rõ yếu tố nào quan trọng nhất.

Ưu điểm:

- Cung cấp giải thích chính xác, dựa trên cơ sở lý thuyết vững chắc.
- Hỗ trợ cả giải thích cục bộ (cho từng dự đoán) và toàn cục (cho toàn bộ mô hình).

Nhược điểm:

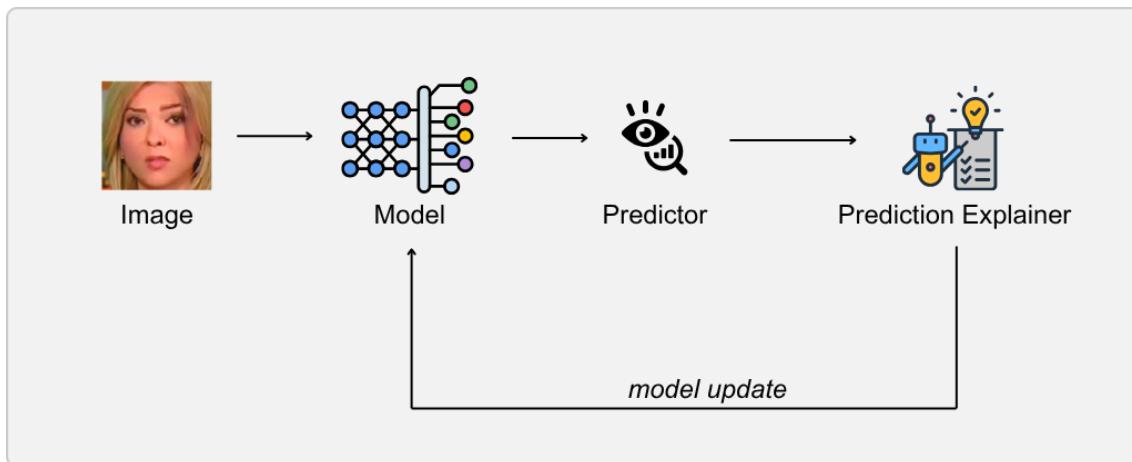
- Tính toán phức tạp, đặc biệt với dữ liệu có nhiều đặc trưng.
- Kết quả có thể khó diễn giải đối với người không quen với lý thuyết trò chơi.

Nhận xét

Các phương pháp XAI truyền thống mặc dù đã giải quyết được vấn đề "hộp đen", tuy nhiên, chúng thường chỉ được sử dụng để phân tích mô hình sau khi huấn luyện, chứ không trực tiếp cải thiện mô hình trong quá trình huấn luyện. Điều này dẫn đến sự ra đời của **Explainability-Guided Approach**, một phương pháp tiên tiến hơn được nhóm đề xuất áp dụng vào phương pháp thực hiện đề tài.

2.5.3 Explainability-Guided Approach

Explainability-Guided Approach là một phương pháp tích hợp khả năng giải thích trực tiếp vào quá trình huấn luyện mô hình học sâu, thay vì chỉ sử dụng XAI để phân tích sau khi huấn luyện. Phương pháp này không chỉ giúp hiểu mô hình mà còn cải thiện hiệu suất của nó trong các bài toán như phát hiện Deepfake.



Hình 2.10: Sơ đồ minh họa quy trình huấn luyện mô hình học sâu với Explainability-Guided Approach, tích hợp Explainer (ví dụ bản đồ giải thích như CAM) để định hướng tập trung vào các đặc trưng khuôn mặt quan trọng trong phát hiện Deepfake.

1. Định nghĩa và mục đích

Định nghĩa: Đây là phương pháp sử dụng các công cụ giải thích (như bản đồ CAM hoặc Grad-CAM) như một tín hiệu hướng dẫn (guidance signal) trong quá trình huấn luyện, nhằm định hướng mô hình học các đặc trưng phù hợp hơn.

Mục đích:

- **Ép mô hình tập trung vào các vùng quan trọng:** Ví dụ, trong phát hiện Deepfake, mô hình sẽ được hướng dẫn để chú ý đến vùng khuôn mặt thay vì các yếu tố không liên quan như nền hoặc nhiễu.
- **Giảm thiểu việc học các đặc trưng không liên quan hoặc nhiễu:** Điều này giúp tránh tình trạng mô hình "học vẹt" các dấu vết giả mạo bề mặt (artifacts) đặc thù của một tập dữ liệu.
- **Tăng khả năng tổng quát hóa và độ tin cậy:** Mô hình trở nên bền vững hơn trước các biến đổi dữ liệu hoặc các cuộc tấn công đối kháng.

2. Cách Thức Hoạt Động

Phương pháp này hoạt động bằng cách tích hợp các bản đồ giải thích vào vòng lặp huấn luyện của mô hình - Hình 2.10. Quy trình cụ thể bao gồm:

1. **Tạo bản đồ giải thích:** Trong mỗi lần huấn luyện, các kỹ thuật như Grad-CAM được sử dụng để sinh ra bản đồ giải thích, chỉ ra các vùng mà mô hình đang tập trung để đưa ra quyết định.
2. **Dánh giá và điều chỉnh:** Nếu bản đồ giải thích cho thấy mô hình chú ý đến các vùng không liên quan (ví dụ: nền thay vì khuôn mặt), tín hiệu từ bản đồ này sẽ được sử dụng để điều chỉnh quá trình huấn luyện.
3. **Hướng dẫn mô hình:** Tín hiệu hướng dẫn (guidance signal) từ bản đồ giải thích được tích hợp vào hàm mất mát (loss function), buộc mô hình phải tập trung vào các vùng quan trọng hơn, chẳng hạn như vùng khuôn mặt trong phát hiện Deepfake.
4. **Lặp lại quá trình:** Quy trình này được thực hiện liên tục trong các vòng huấn luyện, giúp mô hình dần dần học được các đặc trưng có ý nghĩa ngữ nghĩa cao (high-level semantic features).

Ví dụ cụ thể trong phát hiện Deepfake:

- Nếu mô hình ban đầu tập trung vào nhiều nền thay vì khuôn mặt, bản đồ CAM sẽ phát hiện điều này.
- Tín hiệu từ CAM được sử dụng để điều chỉnh trọng số của mô hình, hướng sự chú ý sang các đặc trưng khuôn mặt như mắt, mũi, miệng.
- Kết quả là mô hình không chỉ chính xác hơn mà còn tổng quát hóa tốt hơn trên các tập dữ liệu khác nhau.

2.5.4 Lợi ích của Explainability-Guided Approach

Phương pháp này mang lại nhiều lợi ích vượt trội so với các cách tiếp cận truyền thống:

- **Tăng cường tính bền vững (robustness):** Mô hình ít bị ảnh hưởng bởi các cuộc tấn công đối kháng hoặc các biến đổi nhỏ trong dữ liệu, nhờ vào việc tập trung vào các đặc trưng ngữ nghĩa quan trọng.

- **Cải thiện khả năng tổng quát hóa (generalization):** Thay vì phụ thuộc vào các dấu vết giả mạo đặc thù của một kỹ thuật Deepfake cụ thể, mô hình học được các đặc trưng chung, giúp hoạt động hiệu quả trên nhiều loại dữ liệu.
- **Tăng độ tin cậy và minh bạch:** Việc tích hợp khả năng giải thích vào huấn luyện giúp người dùng và chuyên gia hiểu rõ hơn về quyết định của mô hình, từ đó tăng niềm tin vào hệ thống.

2.5.5 Ứng dụng trong phát hiện Deepfake

Explainability-Guided Approach có thể được áp dụng như một phần của khung Explainability-Guided Adversarial Game, kết hợp với huấn luyện đối kháng để nâng cao hiệu suất phát hiện Deepfake. Cụ thể:

- Mô hình phát hiện (Detector) được huấn luyện cùng với một mô hình tinh chỉnh đối kháng (Adversarial Refiner).
- Bản đồ giải thích (như CAM) được sử dụng để định hướng Detector tập trung vào các đặc trưng khuôn mặt thực sự, thay vì các dấu vết giả mạo bì mặt dễ bị Refiner xóa bỏ.
- Kết quả là Detector không chỉ chính xác hơn mà còn bền vững hơn trước các mẫu Deepfake tinh vi.



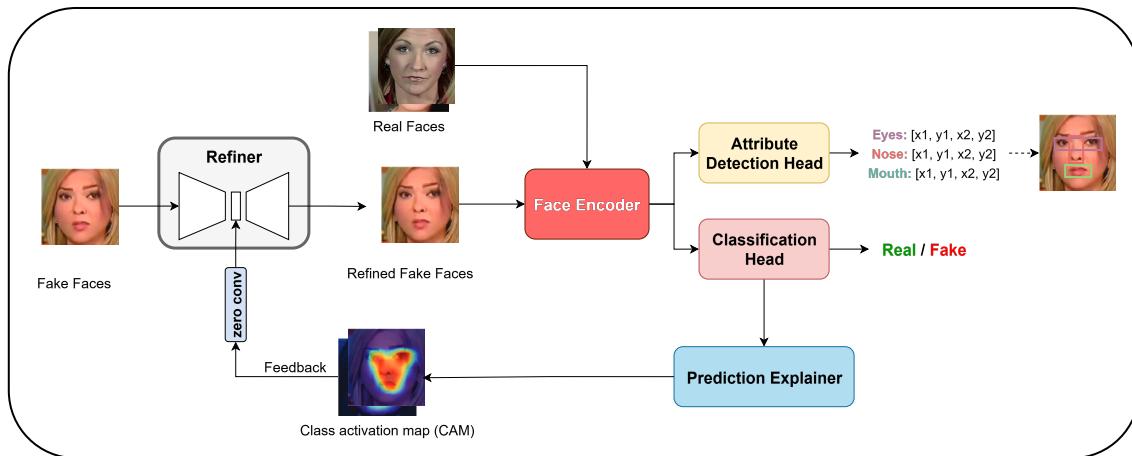
Nhận xét

Explainability-Guided Approach là một phương pháp đột phá, tích hợp khả năng giải thích vào quá trình huấn luyện để cải thiện hiệu suất và độ tin cậy của các mô hình học sâu. Trong bài toán phát hiện Deepfake, nó giúp mô hình tập trung vào các đặc trưng quan trọng, giảm thiểu nhiễu, và tăng khả năng tổng quát hóa. Phương pháp này không chỉ nâng cao tính bền vững của hệ thống mà còn mang lại sự minh bạch, đáp ứng nhu cầu ngày càng cao về độ tin cậy trong các ứng dụng AI thực tế.

CHƯƠNG 3

PHƯƠNG PHÁP THỰC HIỆN

Để giải quyết những thách thức cố hữu về tính bền vững (robustness) và khả năng tổng quát hóa (generalization) của các phương pháp phát hiện hình ảnh giả mạo (Deepfake), chúng tôi đề xuất một khung huấn luyện mới mang tên Trò chơi Đối kháng có Hướng dẫn bởi Khả năng Giải thích (Explainability-Guided Adversarial Game).



Hình 3.1: **Sơ đồ tổng quan kiến trúc Trò chơi Đối kháng có Hướng dẫn bởi Khả năng Giải thích.** Luồng huấn luyện bao gồm: (1) Refiner tinh chỉnh ảnh giả mạo. (2) Detector (Face Encoder và các nhánh đầu ra) phân loại ảnh thật/giả và dự đoán các thành phần khuôn mặt. (3) Prediction Explainer tạo bản đồ CAM từ kết quả phân loại. (4) Bản đồ CAM được sử dụng làm tín hiệu phản hồi để hướng dẫn Refiner trong vòng lặp tiếp theo.

Thay vì phương pháp huấn luyện thụ động truyền thống, chúng tôi thiết lập một hệ sinh thái đối kháng năng động, trong đó một mô hình phát hiện ("người phòng thủ") và một mô hình tinh chỉnh ("kẻ tấn công") cùng cạnh tranh và tiến hóa. Ý tưởng cốt lõi của chúng tôi là tạo ra một vòng lặp phản hồi thông minh và có định hướng. Cụ thể, sau khi mô hình phát hiện phân loại một hình ảnh là giả mạo, chúng tôi sử dụng các kỹ thuật giải thích mô hình (explainability techniques) để tạo ra một "lời giải thích"—một bản đồ nhiệt làm nổi bật các vùng trên ảnh đã dẫn đến quyết định đó. Tín hiệu giải thích này sau đó được phản hồi ngược lại để dẫn dắt mô hình tinh chỉnh. Nhờ đó, mô hình tinh chỉnh không tấn công một cách "mù quáng" mà tập trung

vào việc xóa bỏ hoặc che giấu các dấu vết tại chính những vùng "yếu điểm" mà mô hình phát hiện đã nhận diện. Quá trình này tạo ra các mẫu giả mạo ngày càng tinh vi và khó bị phát hiện hơn.

Trong cuộc chạy đua này, mô hình phát hiện bị buộc phải liên tục thích nghi và nâng cao năng lực. Khi đối mặt với các mẫu tấn công ngày càng thông minh, nó phải từ bỏ việc dựa dẫm vào các dấu vết bề mặt, dễ bị khai thác, và thay vào đó học các đặc trưng bất biến và sâu sắc hơn, từ đó tăng cường đáng kể tính bền vững và khả năng tổng quát hóa. Hơn nữa, để tăng cường nhận thức ngữ nghĩa cho mô hình phát hiện, chúng tôi tích hợp một nhiệm vụ phụ trợ: dự đoán vị trí của các thành phần chính trên khuôn mặt (như mắt, mũi, miệng). Lý do là vì các kỹ thuật Deepfake thường tập trung thao tác trên các vùng này. Bằng cách buộc mô hình phải học các đặc trưng có ý nghĩa về cấu trúc khuôn mặt, chúng tôi không chỉ cung cấp thêm thông tin ngữ nghĩa mà còn giúp mô hình tổng quát hóa tốt hơn, góp phần nâng cao hiệu suất tổng thể.

3.1 Model Overview

3.2 Detector with Multi-task Learning

Để xây dựng một mô hình phát hiện không chỉ chính xác mà còn có khả năng tổng quát hóa cao, chúng tôi đề xuất một mô hình Phát hiện (Detector), ký hiệu là D , được thiết kế dựa trên nguyên lý học đa nhiệm (multi-task learning) [36]. Thay vì chỉ tập trung vào nhiệm vụ phân loại nhị phân (thật/giả), chúng tôi tích hợp một nhiệm vụ phụ trợ nhằm mục đích buộc mô hình phải học các đặc trưng có ý nghĩa ngữ nghĩa về cấu trúc khuôn mặt. Cụ thể, mô hình D được huấn luyện để đồng thời phân biệt ảnh giả mạo và xác định vị trí của các thành phần chính trên khuôn mặt (mắt, mũi và miệng), vốn là những khu vực thường bị tác động mạnh nhất bởi các kỹ thuật giả mạo. Cách tiếp cận này giúp mô hình ít phụ thuộc hơn vào các dấu vết giả mạo bề mặt và thay vào đó học các biểu diễn đặc trưng bền vững hơn.

3.2.1 Architecture

3.2.2 Detector Loss Function

Hàm mất mát tổng thể của Detector, \mathcal{L}_D , là một tổng có trọng số của hai hàm mất mát tương ứng với hai nhiệm vụ, cho phép tối ưu hóa đồng thời cả hai mục tiêu.

1. **Mất mát Phân loại (\mathcal{L}_{cls})**: Đối với nhiệm vụ phân loại nhị phân, chúng tôi sử dụng hàm Binary Cross-Entropy (BCE) tiêu chuẩn, được định nghĩa là:

$$\mathcal{L}_{\text{cls}} = -[y \log(\hat{y}_{\text{cls}}) + (1 - y) \log(1 - \hat{y}_{\text{cls}})] \quad (3.1)$$

với $y \in \{0, 1\}$ là nhãn thật (1 cho ảnh giả, 0 cho ảnh thật).

2. **Mất mát Phát hiện Thành phần ($\mathcal{L}_{\text{comp}}$)**: Đối với nhiệm vụ hồi quy tọa độ hộp giới hạn, chúng tôi sử dụng hàm mất mát **Smooth L1 Loss** [37], một lựa chọn bền vững, kết hợp ưu điểm của L1 Loss (ít nhạy cảm với ngoại lệ) và L2 Loss (tối ưu mượt mà khi lỗi nhỏ). Hàm Smooth L1 được định nghĩa:

$$\text{smooth_L1}(d) = \begin{cases} 0.5d^2 & \text{if } |d| < 1 \\ |d| - 0.5 & \text{otherwise} \end{cases} \quad (3.2)$$

Hàm mất mát tổng thể cho các thành phần được tính bằng cách lấy trung bình Smooth L1 Loss trên tất cả các tọa độ của các hộp giới hạn:

$$\mathcal{L}_{\text{comp}} = \frac{1}{K} \sum_{i=1}^K \sum_{j \in \{x_1, y_1, x_2, y_2\}} \text{smooth_L1}(b_{i,j} - \hat{b}_{i,j}) \quad (3.3)$$

trong đó K là số lượng thành phần (ví dụ $K = 3$ cho mắt, mũi, miệng), $b_{i,j}$ và $\hat{b}_{i,j}$ lần lượt là các giá trị tọa độ thực tế và dự đoán.

Hàm mất mát tổng thể cho Detector được định nghĩa là:

$$\mathcal{L}_D = \mathcal{L}_{\text{cls}} + \lambda_{\text{comp}} * \mathcal{L}_{\text{comp}} \quad (3.4)$$

trong đó λ_{comp} là một siêu tham số để cân bằng tầm quan trọng giữa hai nhiệm vụ.

3.3 Explainability-Guided Adversarial Refiner

Thành phần đối kháng trong trò chơi của chúng tôi là một mô hình Tinh chỉnh (Refiner), ký hiệu là R , có vai trò như một "kẻ tấn công" thông minh. Thay vì tạo ảnh từ đầu, R học cách tinh chỉnh các ảnh giả mạo có sẵn để chúng trở nên khó bị phát hiện hơn. Điểm cốt lõi là quá trình tinh chỉnh này được dẫn dắt bởi tín hiệu giải thích từ Detector, tạo nên một cơ chế tấn công có chủ đích.

3.3.1 Architecture and Refinement Process

Chúng tôi triển khai Refiner R bằng một kiến trúc **Encoder-Decoder** đối xứng. Để đảm bảo Refiner có một khởi đầu tốt, chúng tôi tiến hành tiền huấn luyện nó trên tập ảnh giả mạo với nhiệm vụ tái tạo ảnh (image reconstruction), giúp nó học được các đặc trưng cơ bản của khuôn mặt trước khi tham gia vào quá trình học đối kháng.

Quá trình tinh chỉnh kết hợp thông tin từ bản đồ giải thích M_{cam} (sẽ được mô tả ở mục sau) vào không gian tiềm ẩn của ảnh. Cụ thể, cho một ảnh giả mạo x_{fake} , quá trình tinh chỉnh diễn ra như sau:

1. Ảnh x_{fake} được đưa qua bộ mã hóa của Refiner, R_E , để thu được một biểu diễn không gian tiềm ẩn z_{fake} .
2. Bản đồ giải thích M_{cam} , được tạo ra từ Detector, được đưa qua một lớp tích chập đặc biệt (*zero_conv*) và cộng vào z_{fake} .
3. Biểu diễn tiềm ẩn đã được điều chỉnh $z_{refined}$ được đưa qua bộ giải mã của Refiner, R_D , để tạo ra ảnh tinh chỉnh cuối cùng $x_{refined}$.

Công thức hóa, ta có:

$$z_{fake} = R_E(x_{fake}) \quad (3.5)$$

$$z_{refined} = z_{fake} + \text{zero_conv}(M_{cam}) \quad (3.6)$$

$$x_{refined} = R_D(z_{refined}) \quad (3.7)$$

Ở đây, ‘zero_conv’ là một lớp tích chập có các trọng số được khởi tạo bằng không. Kỹ thuật này, được lấy cảm hứng từ các công trình về điều khiển mạng sinh [38], đảm bảo rằng ở những bước đầu của quá trình huấn luyện, khi tín hiệu từ M_{cam} có thể còn nhiều, ảnh hưởng của nó lên quá trình tinh chỉnh là bằng không. Điều này giúp ổn định quá trình huấn luyện. Sau đó, các trọng số của ‘zero_conv’ sẽ được cập nhật dần thông qua tối ưu hóa, cho phép mô hình học cách tận dụng thông tin từ M_{cam} một cách hiệu quả nhất.

3.3.2 Prediction Explainer (CAM)

Để dẫn dắt Refiner, chúng tôi sử dụng kỹ thuật **Class Activation Map (CAM)** [12], một phương pháp XAI tạo bản đồ nhiệt M_{cam} làm nổi bật vùng ảnh quan trọng nhất cho quyết định phân loại. M_{cam} là ma trận trực quan hóa "sự chú ý" của Detector, tính bằng tổng có trọng số của các bản đồ đặc trưng từ lớp tích chập cuối cùng của bộ mã hóa E_{face} .

$$M_{\text{cam}}(x) = \text{ReLU} \left(\sum_k w_k^c A_k(x) \right) \quad (3.8)$$

trong đó A_k là bản đồ đặc trưng thứ k và w_k^c là trọng số tương ứng với lớp ‘fake’ (c) cho bản đồ đặc trưng thứ k , được lấy từ nhánh phân loại H_c .

3.3.3 Refiner Loss Function

Mục tiêu kép của Refiner R : vừa phải đánh lừa được Detector, vừa phải đảm bảo chất lượng và cấu trúc của ảnh sau khi tinh chỉnh. Do đó, hàm mất mát của nó, \mathcal{L}_R , bao gồm hai thành phần:

- Mất mát Đối kháng (\mathcal{L}_{adv}):** Thành phần này khuyến khích R tạo ra các ảnh mà Detector phân loại là thật. Điều này tương đương với việc tối thiểu hóa xác suất ảnh tinh chỉnh bị phân loại là giả.

$$\mathcal{L}_{\text{adv}} = -\log(1 - D_c(x_{\text{refined}})) \quad (3.9)$$

- Mất mát Tái tạo (\mathcal{L}_{rec}):** Để ngăn Refiner tạo ra các hình ảnh khác biệt quá nhiều so với ảnh gốc (ví dụ như phá hủy cấu trúc khuôn mặt), chúng tôi sử dụng mất mát L1. Thành phần này buộc ảnh tinh chỉnh phải giữ được sự tương đồng về mặt pixel với ảnh giả mạo ban đầu.

$$\mathcal{L}_{\text{rec}} = \|x_{\text{refined}} - x_{\text{fake}}\|_1 \quad (3.10)$$

Hàm mất mát tổng thể cho Refiner là một tổng có trọng số:

$$\mathcal{L}_R = \mathcal{L}_{\text{adv}} + \lambda_{\text{rec}} * \mathcal{L}_{\text{rec}} \quad (3.11)$$

trong đó λ_{rec} là một siêu tham số cân bằng giữa mục tiêu đối kháng và mục tiêu bảo toàn cấu trúc.

3.4 Adversarial Training Framework

Quá trình huấn luyện tổng thể là một trò chơi min-max hai người, trong đó Detector D và Refiner R được tối ưu hóa một cách xen kẽ, theo nguyên lý của huấn luyện đối kháng [1]. Mục tiêu của trò chơi này có thể được biểu diễn như sau:

$$\min_{\theta_D} \max_{\theta_R} \mathcal{L}(\theta_D, \theta_R)$$

Trong thực tế, chúng tôi lặp lại hai bước tối ưu hóa sau:

1. **Cập nhật Detector D :** Giữ cố định các trọng số của Refiner θ_R . Huấn luyện Detector D bằng cách tối thiểu hóa hàm mất mát \mathcal{L}_D (Eq. 3.4) trên một minibatch bao gồm cả ảnh thật x_{real} và ảnh giả đã được tinh chỉnh bởi Refiner, $x_{\text{refined}} = R(x_{\text{fake}}, M_{\text{cam}})$. Bước này giúp Detector học cách phân biệt các mẫu tấn công tinh vi nhất từ Refiner.
2. **Cập nhật Refiner R :** Giữ cố định các trọng số của Detector θ_D . Huấn luyện Refiner R bằng cách tối thiểu hóa hàm mất mát \mathcal{L}_R (Eq. 3.11) trên một minibatch các ảnh giả x_{fake} . Bước này giúp Refiner học cách tạo ra các mẫu tấn công hiệu quả hơn dựa trên trạng thái hiện tại của Detector.

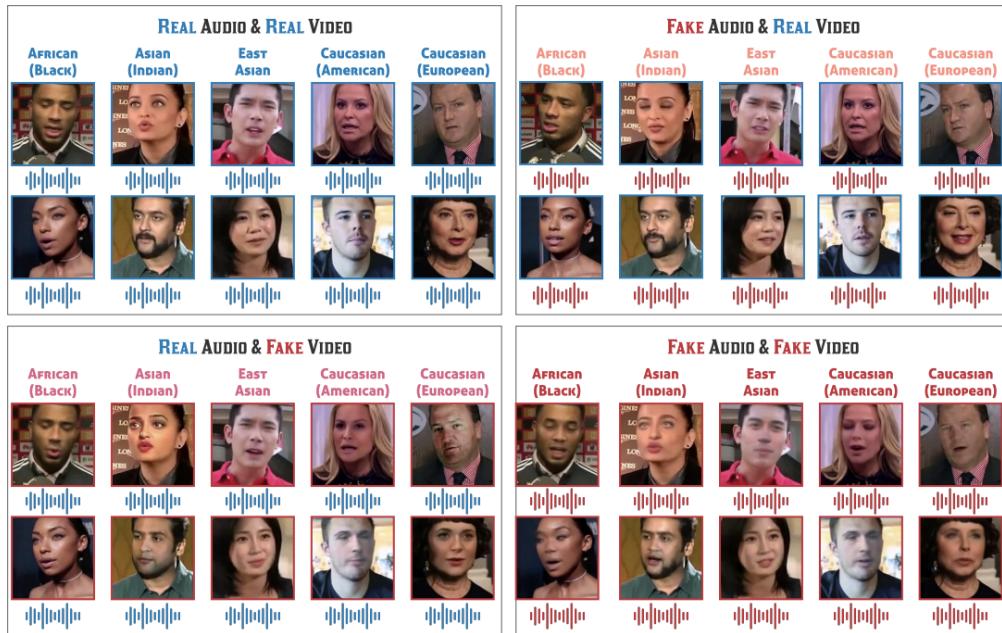
Quá trình cạnh tranh xen kẽ này thúc đẩy cả hai mô hình cùng phát triển, dẫn đến một mô hình Detector cuối cùng có tính bền vững và khả năng tổng quát hóa cao hơn.

CHƯƠNG 4

THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Bộ dữ liệu

Trong quá trình thực nghiệm, nhóm sử dụng tập dữ liệu **FakeAVCeleb** [39], một tập dữ liệu đa phương thức (audio-visual) được thiết kế đặc biệt cho nhiệm vụ phát hiện nội dung deepfake. Tập dữ liệu bao gồm hơn 5000 video của các nhân vật nổi tiếng, được chia thành hai nhóm: video thật và video giả. Các video giả được tạo bằng nhiều kỹ thuật như face swapping, lip-syncing, và audio spoofing, giúp mô phỏng các tình huống giả mạo phức tạp trong thực tế.



Hình 4.1: Hình ảnh minh họa về tập dữ liệu FakeAVCeleb.

FakeAVCeleb cung cấp đầy đủ cả hình ảnh và âm thanh, với nhãn dữ liệu rõ ràng, cho phép mô hình học được đặc trưng giả mạo ở cả hai kênh. Nhờ tính đa dạng và chân thực của dữ liệu, tập này đặc biệt phù hợp để huấn luyện và đánh giá mô hình phát hiện deepfake kết hợp GAN và khả năng giải thích, đồng thời kiểm tra độ tin cậy của mô hình trong các tình huống giả mạo đa dạng và khó phát hiện hơn.

4.2 Độ đo

Để đánh giá hiệu suất của mô hình trong bài toán phát hiện và phân loại hình ảnh Deepfake, chúng tôi sử dụng độ đo **Accuracy** (**Độ chính xác**). Đây là một trong những độ đo phổ biến và trực quan nhất để đánh giá hiệu suất tổng thể của một mô hình phân loại nhị phân.

Độ chính xác được định nghĩa là tỷ lệ giữa số lượng dự đoán đúng trên tổng số dự đoán được thực hiện. Công thức tính toán như sau:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Trong đó:

- **TP (True Positives)**: Số lượng hình ảnh giả mạo (fake) được mô hình dự đoán chính xác là giả mạo.
- **TN (True Negatives)**: Số lượng hình ảnh thật (real) được mô hình dự đoán chính xác là thật.
- **FP (False Positives)**: Số lượng hình ảnh thật bị mô hình dự đoán nhầm là giả mạo.
- **FN (False Negatives)**: Số lượng hình ảnh giả mạo bị mô hình dự đoán nhầm là thật.

Trong bối cảnh của nghiên cứu này, Accuracy cho biết khả năng của mô hình trong việc phân biệt chính xác giữa hình ảnh thật và hình ảnh đã qua chỉnh sửa bằng công nghệ Deepfake. Một giá trị Accuracy cao thể hiện rằng mô hình có hiệu suất tốt trên cả hai lớp và ít mắc phải sai lầm trong việc phân loại.

4.3 Cài đặt thực nghiệm

Tiền xử lý dữ liệu. Đầu tiên, chúng tôi trích xuất 32 khung hình từ mỗi video và áp dụng mô hình 2D-FAN [40]¹ để phát hiện 68 điểm mốc trên khuôn mặt. Các điểm mốc này sau đó được sử dụng để thực hiện một phép biến đổi tương tự, nhằm căn chỉnh và chuẩn hóa các khuôn mặt về vị trí, tỷ lệ và góc xoay, trước khi cắt ra với một kích thước cố định. Cuối cùng, chúng tôi tận dụng chính tập điểm mốc này để tự động tạo ra các hộp giới hạn

¹<https://github.com/1adrianb/face-alignment>

(bounding boxes) cho các vùng mắt, mũi, và miệng, cung cấp dữ liệu nhãn (ground truth) cho nhiệm vụ học đa nhiệm của mô hình.

Chi tiết thực nghiệm. Đối với bộ mã hóa đặc trưng (Face Encoder), chúng tôi lựa chọn kiến trúc ResNet-18 [23], được khởi tạo với các trọng số đã được huấn luyện trước trên tập dữ liệu ImageNet [41] để tăng tốc độ hội tụ và cải thiện hiệu suất. Mô hình được huấn luyện trong 5 epochs với batch size là 16. Chúng tôi sử dụng bộ tối ưu AdamW [42] với tốc độ học ban đầu (initial learning rate) được thiết lập là 1×10^{-4} và hệ số suy giảm trọng số (weight decay) là 1×10^{-3} để tối ưu hóa mô hình. Hiệu suất của phương pháp đề xuất được đánh giá thông qua thước đo Độ chính xác (Accuracy) trên tập dữ liệu kiểm tra.

4.3.1 So sánh với các Phương pháp khác

Bảng 4.1 trình bày kết quả so sánh hiệu suất giữa phương pháp của chúng tôi (được gọi là EAG - Explainability-Guided Adversarial Game) và các phương pháp hiện đại khác. Phương pháp của chúng tôi đạt độ chính xác là **95.10%**, một kết quả rất cạnh tranh trong bối cảnh của một tập dữ liệu phức tạp như FakeAVCeleb.

Bảng 4.1: So sánh hiệu suất của các nghiên cứu trên tập dữ liệu FakeAVCeleb.

Phương pháp	Accuracy (%)
Multimodaltrace [43]	92.90
AVFakeNet [44]	93.40
AV-Lip-Sync [45]	94.00
PVASS-MDD [46]	95.70
MIS-AViODD [47]	96.70
Ours	95.10

Phân tích kết quả cho thấy, mặc dù phương pháp của chúng tôi có độ chính xác thấp hơn một chút so với MIS-AViODD (96.70%) và PVASS-MDD (95.70%), nó vẫn vượt trội hơn đáng kể so với các phương pháp khác như AV-Lip-Sync, AVFakeNet và Multimodaltrace. Điều đáng chú ý là nhiều phương pháp hàng đầu như MIS-AViODD thường tận dụng cả hai luồng dữ liệu âm thanh (audio) và hình ảnh (visual) để phát hiện sự không đồng nhất. Ngược lại, phương pháp của chúng tôi, mặc dù chỉ tập trung vào phân tích hình ảnh, đã chứng tỏ tiềm năng mạnh mẽ của cơ chế trò chơi đối kháng có hướng dẫn. Kết quả này khẳng định rằng việc buộc mô hình phát hiện học các đặc trưng bền vững thông qua cạnh tranh với một tác nhân tinh chỉnh thông minh là

một hướng đi hiệu quả và có thể sánh ngang với các phương pháp đa phương thức phức tạp.

4.3.2 Nghiên cứu Tách biệt (Ablation Study)

Để đánh giá tầm quan trọng và sự đóng góp của từng thành phần cốt lõi trong kiến trúc đề xuất, chúng tôi đã tiến hành một nghiên cứu tách biệt. Chúng tôi lần lượt loại bỏ hai thành phần chính: (1) Mô hình Tinh chỉnh Đối kháng (*Adversarial Refiner*), tương đương với việc chỉ huấn luyện mô hình phát hiện một cách thông thường; và (2) Nhánh Phát hiện Thành phần (*Component Head*), để xem xét tác động của việc học đa nhiệm. Kết quả được trình bày chi tiết trong Bảng 4.2.

Bảng 4.2: Kết quả nghiên cứu tách biệt, thể hiện sự ảnh hưởng của các thành phần trong mô hình trên tập dữ liệu FakeAVCeleb.

Cấu hình Mô hình	Accuracy (%)
Ours w/o Adversarial Refiner	94.09
Ours w/o Component Head	94.42
Ours (Full Model)	95.10

Từ bảng kết quả, chúng ta có thể rút ra các nhận xét quan trọng sau:

- **Tầm quan trọng của Trò chơi Đối kháng:** Khi loại bỏ toàn bộ cơ chế Tinh chỉnh Đối kháng ('w/o Adversarial Refiner'), hiệu suất của mô hình giảm mạnh nhất, xuống còn 94.09% (giảm 1.01% so với mô hình đầy đủ). Điều này khẳng định rằng trò chơi đối kháng có hướng dẫn bởi khả năng giải thích là nhân tố quan trọng nhất, đóng vai trò then chốt trong việc thúc đẩy mô hình phát hiện học các đặc trưng bền vững và có khả năng tổng quát hóa cao hơn.
- **Đóng góp của Học Đa nhiệm:** Tiếp theo, khi chúng tôi loại bỏ Nhánh Phát hiện Thành phần ('w/o Component Head'), hiệu suất cũng giảm xuống 94.42% (giảm 0.68%). Sự sụt giảm này chứng tỏ rằng việc buộc mô hình học các đặc trưng ngữ nghĩa về cấu trúc khuôn mặt thông qua một nhiệm vụ phụ trợ thực sự góp phần cải thiện khả năng phân loại cuối cùng, giúp mô hình tập trung vào các vùng có ý nghĩa thay vì các dấu vết giả mạo ngẫu nhiên.

Khi kết hợp đầy đủ cả hai thành phần, mô hình của chúng tôi đạt hiệu suất cao nhất là 95.10%. Tổng hợp lại, các kết quả từ nghiên cứu tách biệt

đã chứng minh một cách thuyết phục rằng cả hai đóng góp chính của chúng tôi - cơ chế trò chơi đối kháng có hướng dẫn và kiến trúc học đa nhiệm - đều có tác động tích cực và bổ trợ cho nhau để tạo nên hiệu quả vượt trội của mô hình cuối cùng.

CHƯƠNG 5

KẾT LUẬN

Nghiên cứu này đã đề xuất và triển khai thành công một phương pháp mới mang tên "Trò chơi Đối kháng có Hướng dẫn bởi Khả năng Giải thích" (Explainability-Guided Adversarial Game), nhằm giải quyết thách thức cốt lõi về độ bền vững và khả năng tổng quát hóa của các trình phát hiện hình ảnh Deepfake. Thay vì các phương pháp huấn luyện truyền thống, chúng tôi đã xây dựng một khuôn khổ đối kháng năng động, nơi mô hình phát hiện (Detector) và mô hình tinh chỉnh (Refiner) cạnh tranh và cùng tiến hóa.

5.1 Điểm mạnh

Điểm đột phá của phương pháp nằm ở vòng lặp phản hồi thông minh, trong đó Refiner không tấn công một cách "mù quáng" mà được dẫn dắt trực tiếp bởi các bản đồ giải thích (CAM) từ Detector. Cơ chế này buộc Detector phải từ bỏ sự phụ thuộc vào các dấu vết giả mạo bề mặt, dễ bị loại bỏ, và thay vào đó, học các đặc trưng ngữ nghĩa sâu sắc và bền vững hơn về cấu trúc khuôn mặt. Hơn nữa, việc tích hợp nhiệm vụ học đa nhiệm—yêu cầu mô hình đồng thời xác định vị trí các thành phần quan trọng như mắt, mũi, và miệng—đã tăng cường đáng kể khả năng nhận thức về ngữ nghĩa, giúp mô hình tập trung vào các vùng thao tác trọng yếu.

Kết quả thực nghiệm trên tập dữ liệu phức tạp FakeAVCeleb đã chứng minh hiệu quả vượt trội của phương pháp, đạt độ chính xác 95.10%. Kết quả này không chỉ cạnh tranh với các phương pháp đa phương thức hàng đầu mà còn khẳng định sức mạnh của việc chỉ tập trung vào phân tích hình ảnh khi được trang bị một cơ chế huấn luyện thông minh. Nghiên cứu tách biệt cũng cho thấy cả hai thành phần cốt lõi—trò chơi đối kháng có hướng dẫn và học đa nhiệm—đều đóng góp một cách tích cực và bổ trợ lẫn nhau để tạo nên hiệu suất cuối cùng.

5.2 Hạn chế

Mặc dù đạt được những kết quả đáng khích lệ, phương pháp của chúng tôi vẫn tồn tại một số hạn chế. Thứ nhất, mô hình hiện tại chỉ tập trung vào

việc phân tích dữ liệu hình ảnh (visual). Điều này đồng nghĩa với việc chúng tôi chưa khai thác các thông tin hữu ích từ luồng âm thanh (audio), vốn là một nguồn dữ liệu quan trọng trong việc phát hiện các video Deepfake, đặc biệt là các kỹ thuật giả mạo lip-sync hoặc hoán đổi giọng nói. Thứ hai, hiệu suất của mô hình phụ thuộc vào chất lượng của các bản đồ giải thích được tạo ra. Nếu các bản đồ này nhiễu hoặc không chính xác, chúng có thể định hướng sai cho quá trình tinh chỉnh của Refiner, từ đó ảnh hưởng đến hiệu quả chung của cả hệ thống. Cuối cùng, độ phức tạp của khung huấn luyện đối kháng đòi hỏi chi phí tính toán cao hơn so với việc huấn luyện một mô hình phát hiện đơn lẻ.

5.3 Hướng phát triển trong tương lai

Dựa trên những kết quả đã đạt được và các hạn chế còn tồn tại, chúng tôi đề xuất một số hướng phát triển tiềm năng trong tương lai:

- **Tích hợp đa phương thức (Multimodality):** Mở rộng kiến trúc để kết hợp cả luồng dữ liệu âm thanh và hình ảnh. Việc phân tích sự đồng bộ và nhất quán giữa hai kênh thông tin này hứa hẹn sẽ tạo ra một hệ thống phát hiện mạnh mẽ và toàn diện hơn, có khả năng đối phó với nhiều kỹ thuật giả mạo tinh vi hơn.
- **Cải tiến cơ chế giải thích:** Nghiên cứu và áp dụng các kỹ thuật XAI tiên tiến hơn để tạo ra các bản đồ giải thích chi tiết và chính xác hơn. Điều này sẽ giúp cung cấp tín hiệu phản hồi chất lượng hơn cho Refiner, thúc đẩy cuộc cạnh tranh đối kháng lên một tầm cao mới.
- **Khám phá các kiến trúc đối kháng mới:** Thử nghiệm các kiến trúc Refiner và Detector khác nhau, cũng như các hàm mất mát mới, để tối ưu hóa quá trình học đối kháng và cải thiện hơn nữa khả năng tổng quát hóa của mô hình trước các kỹ thuật Deepfake chưa từng gặp.

Tóm lại, nghiên cứu này đã chứng tỏ rằng việc kết hợp huấn luyện đối kháng với khả năng giải thích là một hướng đi đầy hứa hẹn để xây dựng các thế hệ trình phát hiện Deepfake bền vững và đáng tin cậy hơn, góp phần quan trọng vào việc bảo vệ sự toàn vẹn của thông tin trong không gian số.

TÀI LIỆU THAM KHẢO

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in neural information processing systems*, vol. 27, 2014.
- [2] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [3] OpenAI, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [4] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
- [5] R. Chesney and D. Citron, “Deep fakes: A looming challenge for privacy, democracy, and national security,” *Available at SSRN 3213954*, 2019.
- [6] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A survey of face manipulation and fake detection,” in *Information Fusion*, Elsevier, vol. 64, 2020, pp. 131–148.
- [7] M. Westerlund, “The rise of deepfakes: A new frontier for corporate and criminal liability,” *Computer Law & Security Review*, vol. 36, p. 105 373, 2020.
- [8] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: A compact facial video forgery detection network,” in *2018 IEEE international workshop on information forensics and security (WIFS)*, IEEE, 2018, pp. 1–7.
- [9] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, “Leveraging frequency analysis for deep fake image recognition,” in *International conference on machine learning*, PMLR, 2020, pp. 3247–3258.
- [10] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in *2018 IEEE international workshop on information forensics and security (WIFS)*, IEEE, 2018, pp. 1–7.

- [11] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, “Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3348–3357.
- [12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
- [13] R. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: Cnn based generative deep neural networks are failing to produce high-quality an-aliased images,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8090–8099.
- [14] U. A. Ciftci, I. Demir, and L. Yin, “How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals,” in *2020 IEEE international joint conference on biometrics (IJCB)*, IEEE, 2020, pp. 1–10.
- [15] C. Peng, Z. Miao, D. Liu, N. Wang, R. Hu, and X. Gao, “Where deep-fakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4507–4520, 2024.
- [16] Z. Wang, Y. Guo, and W. Zuo, “Deepfake forensics via an adversarial game,” *IEEE Transactions on Image Processing*, vol. 31, pp. 3541–3552, 2022.
- [17] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in *International Conference on Learning Representations*, 2018.
- [18] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.
- [19] Y. Nirkin, Y. Keller, and T. Hassner, “Fsgan: Subject agnostic face swapping and reenactment,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.

-
- [20] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.
 - [21] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, “Attgan: Facial attribute editing by only changing what you want,” in *IEEE transactions on image processing*, vol. 28, IEEE, 2019, pp. 5464–5478.
 - [22] M. Liu, Y. Liu, Q. Li, Y. Li, C. Shen, and M. Tang, “Stgan: A unified selective transfer network for arbitrary image attribute editing,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3673–3682.
 - [23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
 - [24] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.
 - [25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
 - [26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *arXiv preprint arXiv:1704.04861*, 2017.
 - [27] L. Chai, D. Bau, S.-N. Lim, and P. Isola, “What makes fake images detectable? a patch-based analysis,” in *European conference on computer vision*, Springer, 2020, pp. 616–632.
 - [28] H. Li, Z. Wang, and K. Li, “Dcta: A dct-based attention module for face forgery detection,” in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2020, pp. 1–6.
 - [29] H. Li, Z. Wang, and K. Li, “Dftd: A dft-based transformer for deepfake detection,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 2605–2609.
 - [30] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.

-
- [31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
 - [32] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *Workshop at International Conference on Learning Representations*, 2014.
 - [33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.
 - [34] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
 - [35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, vol. 30, 2017.
 - [36] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
 - [37] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
 - [38] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," in *IEEE transactions on pattern analysis and machine intelligence*, IEEE, vol. 41, 2018, pp. 1947–1962.
 - [39] H. Khalid, S. Shah, M. Wahab, M. Khan, M. M. Fraz, Z. Z-Ul-Abidin, and M. U. Akram, "Fakeavceleb: A novel audio-video deepfake dataset," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 3343–3347.
 - [40] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1021–1030.

-
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255.
 - [42] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
 - [43] R. Zhou, J. Han, and J. Zhang, “Multimodaltrace: A multimodal-based method for deepfake detection,” in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2021, pp. 1–6.
 - [44] K. Li, J. Zhang, and Z. Wang, “Avfakenet: A two-stream audiovisual network for deepfake video detection,” in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 2610–2614.
 - [45] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.
 - [46] Y. Chen, H. Li, and Z. Wang, “Pvass-mdd: A privacy-preserving and verifiable audio-visual splicing-based method for deepfake detection,” in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 1–6.
 - [47] Y. Wu, H. Li, and Z. Wang, “Mis-aviodd: Modality-independent and -specific representation learning for audio-visual deepfake detection,” in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
 - [48] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.