HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

PHẠM THANH HÙNG

PHÁT HIỆN GIẢ MẠO KHUÔN MẶT SỬ DỤNG MẠNG HỌC SÂU

LUẬN VĂN THẠC SĨ KỸ THUẬT

(Theo định hướng ứng dụng)

HÀ NỘI - 2022

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

PHẠM THANH HÙNG

PHÁT HIỆN GIẢ MẠO KHUÔN MẶT SỬ DỤNG MẠNG HỌC SÂU

Chuyên ngành : **KHOA HỌC MÁY TÍNH**

Mã số: 8.48.01.01

LUẬN VĂN THẠC SĨ KỸ THUẬT

NGƯỜI HƯỚNG DẪN KHOA HỌC

GS. TS. TỪ MINH PHƯƠNG

HÀ NỘI - 2022

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu và tìm hiểu của riêng tôi.

Các số liệu, kết quả nêu trong luận văn là trung thực và chưa từng được ai công bố trong bất cứ công trình nào khác.

Tác giả luận văn

Phạm Thanh Hùng

LỜI CẢM ƠN

Để có thể hoàn thành được luận văn này, em xin gửi lời cảm ơn sâu sắc tới thầy Từ Minh Phương, người đã trực tiếp hướng dẫn, tận tình chỉ bảo và đưa ra cho em những lời khuyên cho em trong suốt quá trình nghiên cứu và thực hiện luận văn. Không những thế, trong cuộc sống, thầy cũng được giúp đỡ em rất nhiều để có thể tiếp tục vượt qua những giai đoạn khó khăn trong cuộc sống.

Em cũng xin gửi lời cảm ơn chân thành cảm ơn tất cả các thầy cô giáo của Học viện Công nghệ Bưu chính Viễn thông đã giảng dạy và dìu dắt em trong suốt quá trình học tập tại trường.

Cuối cùng, em xin gửi lời cảm ơn tới gia đình, bạn bè và những người đã luôn ở bên cổ vũ tinh thần, tạo điều kiện thuận lợi cho em để em có thể học tập tốt và hoàn thiện luận văn.

Dù đã cố gắng hết sức trong quá trình làm luận văn nhưng cũng không thể tránh khỏi những sai sót, em mong nhận được sự thông cảm và đóng góp ý kiến của các thầy cô để luận văn có thể được hoàn thiện tốt hơn nữa!

Em xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC CHỮ CÁI VIẾT TẮT

Từ viết tắt	Tiếng Anh	Tiếng Việt
CNN	Convolutional neural network	Mạng tích chập
SURF	speed-up robust features	
CTMF	Colour Texture Markov feature	Đặc trưng cấu trúc màu Markov
SMV-RFE	Support Vector Machine-recursive Feature Elimination	Máy vector hỗ trợ loại bỏ đệ quy
OFL	Optical Flow Line	
rPPG	Remote Photoplethysmography	
CFrPPG	rPPG correspondence feature	
IDA	Image distortion analysis	Phân tích biến dạng của ảnh
RNN	Recursive neural network	Mạng nơ-ron hồi quy
OFFB	Optical Flow guided Feature Block	
LSTM	Long short-term memory	

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

MỞ ĐẦU

Ngày nay, các ứng dụng của trí tuệ nhân tạo ngày càng trở lên phổ biến, một trong các ứng dụng đó là nhận diện khuôn mặt. Tuy nhiên, sự phát triển của các ứng dụng này cũng kéo theo một vấn đề đó là phát hiện giả mạo khuôn mặt. Thuật ngữ 'giả mạo khuôn mặt' ở đây nhằm nói đến việc xây dựng các khuôn mặt giả mạo của một người thật bằng nhiều cách khác nhau như lấy ảnh chụp của người đó in ra giấy hay quay lại được một video có khuôn mặt của họ. Tất cả các hành động trên nhằm đánh lừa hệ thống nhận diện khuôn mặt rằng nạn nhân đang có mặt tại thời điểm xác thực khuôn mặt, từ đó đạt được các mục đích xấu như vượt qua các biện pháp bảo mật nhằm đánh cắp tiền hay đánh cắp thông tin cá nhân v.v. Thật vậy, trong những năm gần đây, xu hướng 4.0 trở thành mối quan tâm lớn của toàn xã hội khi sự phát triển vượt bậc và nhanh chóng của ngành công nghệ thông tin nói chung và trí tuệ nhân tạo nói riêng đã mang lại những tiện lợi vô cùng to lớn cho con người. Cụ thể với bài toán nhận dạng khuôn mặt, thuật toán này có đầu vào là một hình ảnh có chứa khuôn mặt của một người nào đó, thuật toán sẽ đưa ra thông tin người đó là ai dựa trên một tập cơ sở dữ liệu khuôn mặt đã được thu thập từ trước. Với khả năng như vậy, nhận diện khuôn mặt đã được áp dụng vào nhiều lĩnh vực như an ninh tại sân bay, với mục tiêu phát hiện các phần tử khủng bố, tăng sự an toàn cho ngành hàng không. Trong môi trường doanh nghiệp, nhận diện khuôn mặt được sử dụng để kiểm soát ra vào, chấm công tự động, bảo mật máy tính. Hay trong lĩnh vực ngân hàng, đó là eKYC. Xác thực người dùng khi đăng ký sử dụng các dịch vụ của ngân hàng dựa vào hình ảnh mà người dùng chụp trực tiếp từ điện thoại khi đăng ký. Như vậy, các hệ thống nhận diện khuôn mặt này đóng vai trò vô cùng quan trọng trong các nghiệp vụ đã được liệt kê phía trên. Bởi khi các hệ thống này bị vượt qua thì sẽ để lại những thiệt hại vô cùng lớn cho người dùng và các doanh nghiệp khi thông tin của họ có thể bị truy cập trái phép bởi kẻ xấu. Chính vì vậy việc phát triển các giải pháp nhằm chống giả mạo khuôn mặt trong các hệ thống nhận diện là vô cùng quan trọng. Thêm vào đó, sự thành công đáng kinh ngạc của mạng nơ-ron tích chập (convolution neural network - CNN) trong cuộc thi ImageNet [59] đã thu hút rất nhiều sự chú ý của các nhà nghiên cứu trong mảng thị giác máy tính nhằm

khai thác các khả năng tiềm ẩn của phương pháp học sâu. Sự cải tiến ngày càng tăng của mạng CNN nói chung về phân loại hình ảnh và phát hiện đối tượng đã mở ra các nhánh và ứng dụng tiềm năng của CNN trong lĩnh vực chống giả mạo khuôn mặt. Với các lý do trên, em đã chọn đề tài luận văn là "Phát hiện giả mạo khuôn mặt sử dụng mạng học sâu".

Nội dung luận văn được chia thành 3 chương như sau:

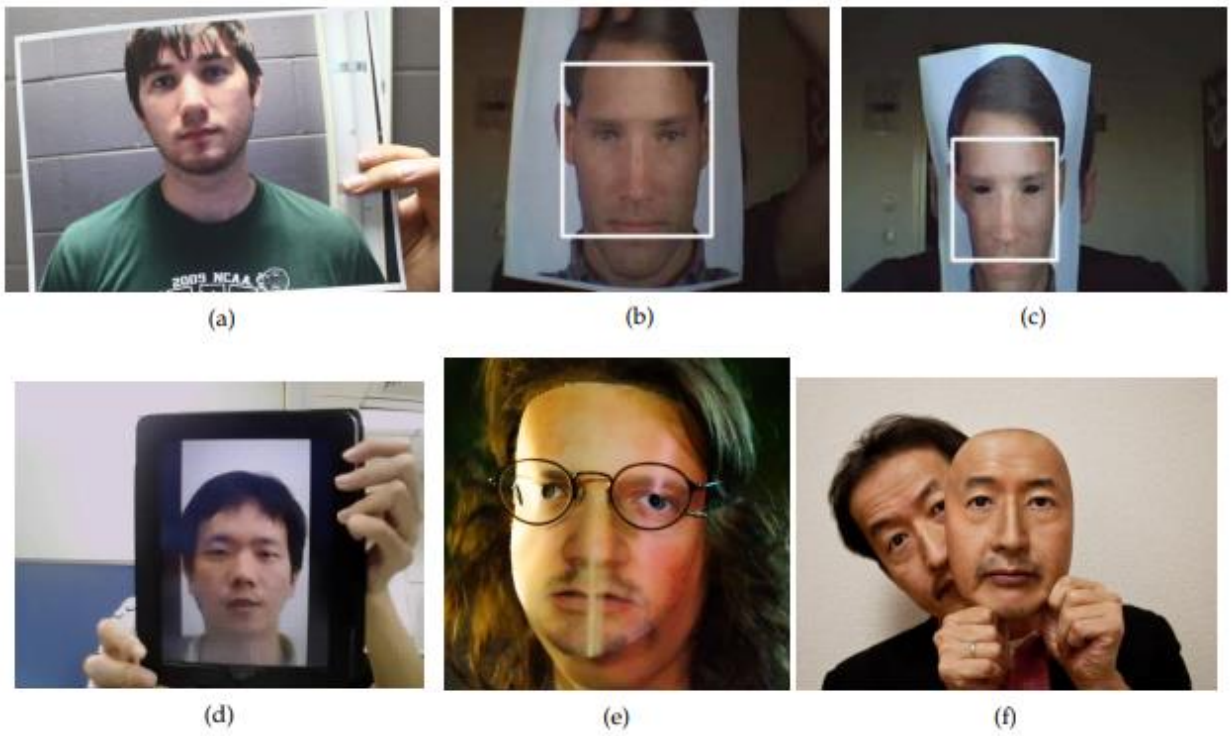
- CHƯƠNG 1: Bài toán phát hiện giả mạo khuôn mặt: Giới thiệu bài toán mà luận văn nghiên cứu và các nghiên cứu liên quan đã có.
- CHƯƠNG 2: Ứng dụng mạng học sâu vào bài toán phát hiện giả mạo khuôn mặt: Đưa ra một số lý thuyết về mạng học sâu, ý tưởng của việc đưa đặc trưng LBP vào mạng tích chập, cách tạo ảnh chiều sâu khuôn mặt từ mạng học sâu, giới thiệu mạng resnet-34, cách kết hợp các kỹ thuật trên. Bên cạnh đó, chương này sẽ nêu ra vấn đề thích ứng miền và ý tưởng sử dụng GAN để hạn chế vấn đề này.
- CHƯƠNG 3: Thử nghiệm và đánh giá: Trình bày về tập dữ liệu, các độ đo, các thử nghiệm, đưa ra các kết quả và rút ra kết luận.

CHƯƠNG 1: BÀI TOÁN PHÁT HIỆN GIẢ MẠO KHUÔN MẶT

Chương này sẽ trình bày định nghĩa của bài toán phát hiện giả mạo khuôn mặt cùng với các nghiên cứu liên quan tới bài toán này. Cụ thể, chương 1 sẽ giới thiệu các phương pháp dựa trên đặc trưng texture, các phương pháp dựa trên tương tác giữa người và máy, các thông tin về sự sống, chất lượng và chiều sâu của hình ảnh cũng được đề cập. Cuối cùng là các phương pháp dựa trên học sâu.

1.1. Giới thiệu bài toán phát hiện giả mạo khuôn mặt

Phát hiện giả mạo khuôn mặt là nhiệm vụ phát hiện hành vi xác minh khuôn mặt bằng cách sử dụng ảnh, video, mặt nạ hoặc một vật thay thế khác cho khuôn mặt của một người.



Hình 1-1: Các phương thức giả mạo khuôn mặt

Dưới đây là một số hình thức giả mạo khuôn mặt hay được sử dụng nhất:

- Giả mạo bằng hình thức in: Kẻ tấn công sử dụng một bức ảnh của nạn nhân, sau đó in ra hoặc hiển thị trên một thiết bị điện tử. Đây là hình thức giả mạo phổ biến nhất do hầu hết ảnh chụp của các cá nhân đều sẵn có trên mạng và có thể lấy đi mà không cần sự cho phép của chủ nhân bức ảnh đó.

- Giả mạo bằng ảnh đục lỗ ở mắt: Vùng mắt của bức ảnh in sẽ được đục lỗ để giả mạo thêm được hành vi chớp mắt của một người.
- Giả mạo bởi ảnh làm cong: Kẻ tấn công sẽ uốn ảnh với nhiều hướng khác nhau để giả mạo biểu cảm của khuôn mặt.
- Giả mạo bằng video: Ở hình thức này, kẻ tấn công đã lấy được một video quay lại khuôn mặt của nạn nhân. Cách này khiến cho hành vi và chuyển động của khuôn mặt giả mạo trông tự nhiên hơn khi có thể có đầy đủ dấu hiệu của sự sống như chớp mắt, nét mặt, chuyển động ở đầu và miệng, cuối cùng là phương thức này dễ dàng được thực hiện bằng máy tính bảng và điện thoại thông minh cỡ lớn.
- Giả mạo bằng mặt nạ 3D: Một chiếc mặt nạ 3D sẽ được sử dụng làm công cụ giả mạo ở hình thức này. Thậm chí giả mạo bằng mặt nạ 3D còn tinh vi hơn cả việc sử dụng video khi có hình chuyển động ở khuôn mặt rất tự nhiên và có thể vượt qua được các thiết bị đặc biệt như cảm biến chiều sâu.

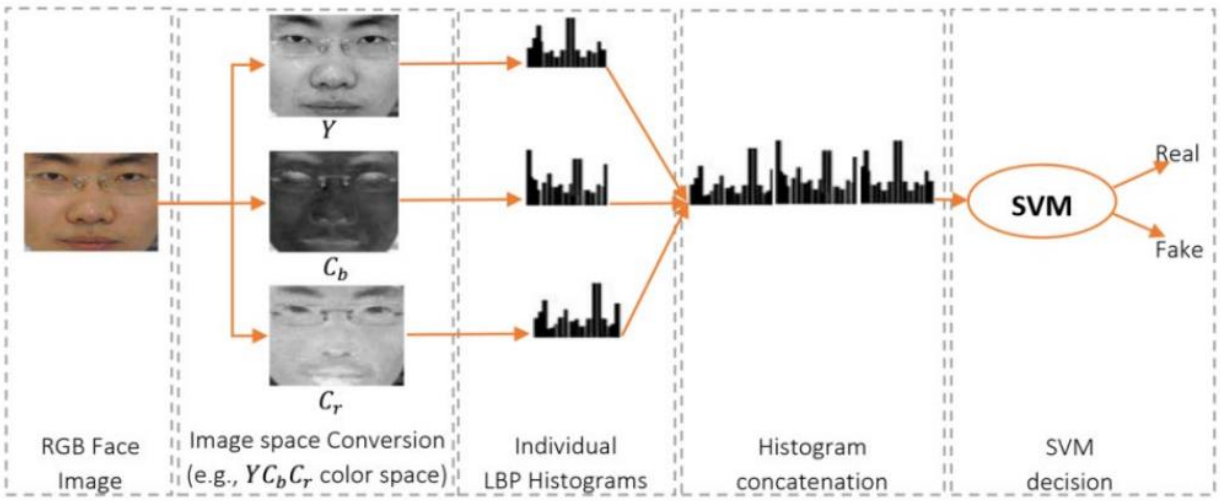
1.2. Các nghiên cứu liên quan

1.2.1. Các phương pháp dựa trên đặc trưng texture của ảnh

Trong quá trình thu nhận ảnh, hình ảnh bị mất mát đi một số thông tin, cùng với đó có một số nhiễu xuất hiện kèm theo quá trình đó. Bên cạnh đó, hình ảnh được thu nhận qua hai lần cũng sẽ có sự khác biệt về mặt kết cấu (texture) nhất định so với ảnh được thu nhận trực tiếp. Sự khác biệt nhỏ này xuất hiện bởi các yếu tố chủ yếu như điểm sáng cục bộ, sự thay đổi bóng và mức độ nhòe mờ của hình ảnh [4]. Các lý do này sẽ là yếu tố chủ đạo để các phương pháp dựa trên kết cấu phân biệt ảnh thật và ảnh giả.

Một số phương pháp thực hiện chuyển đổi ảnh thu thập được sang dạng xám và trích chọn các thông tin về kết cấu trên các ảnh xám này cho bài toán phân loại ảnh mặt thật và ảnh giả. Các phương pháp này bỏ qua thông tin về kết cấu màu của hình ảnh. Maatta và công sự [5] đã sử dụng nhiều toán tử LBP đồng nhất với nhiều tỷ lệ khác nhau để trích xuất histogram đặc trưng về kết cấu từ các vùng địa phương của ảnh xám và ảnh toàn cục. Sau

đó, họ kết nối chúng lại để tạo thành một biểu đồ đặc trưng có 531 chiều và đưa qua bộ phân loại SVM với RBF là lõi cho quá trình huấn luyện và kiểm thử phân loại mặt thật và mặt giả mạo. Các thuật toán phân tích kết cấu dựa trên bản đồ tỷ lệ xám tỏ ra hiệu quả với các ảnh khuôn mặt giả mạo có kết cấu rõ ràng và có độ phân giải cao. Tuy nhiên, đối với các hình ảnh giả mạo có độ phân giải thấp thì việc phân biệt được trở lên khó hơn. Trong khi đó, Boulkenaf và công sự [1] đề xuất một phương pháp phát hiện ảnh giả mạo dựa trên phân tích cấu trúc màu. Cụ thể, Họ trích xuất các biểu đồ histograms LBP từ mỗi kênh màu, sau đó kết nối chúng lại để tạo thành một đặc trưng mô tả cuối cùng. Tiến trình đặc biệt này được mô tả qua hình 1-2. Để phân tích không gian màu nào giúp dễ phân biệt mặt thật và mặt giả mạo hơn, phương pháp này thí nghiệm ba không gian màu RGB, YCbCr và HSV. Các thí nghiệm cho thấy rằng phương pháp dựa trên kết cấu màu sắc vượt trội hơn so với phương pháp dựa trên kết cấu màu xám trong việc phát hiện các cách thức giả mạo khác nhau. Boulkenaf và cộng sự. [1] cũng tập trung vào các kênh độ chói và sắc độ, đồng thời kết hợp các đặc trưng LBP nhiều cấp của khuôn mặt người trong không gian HSV với các đặc trưng LPQ của khuôn mặt người trong không gian YCbCr bằng cách sử dụng thông tin chung về màu sắc và kết cấu. Mặc dù đã đạt được kết quả tốt trong thử nghiệm, nhưng các bộ mô tả kết cấu vi mô ở mức thấp khiến chúng nhạy cảm với sự thay đổi ánh sáng và hình ảnh chất lượng cao. Để cải thiện khả năng phân biệt hơn nữa, Boulkenaf và cộng sự [6] với các đặc tính ổn định được tăng tốc (speed-up robust features, SURF) để đối mặt với việc phát hiện giả mạo. So với các phương pháp trước đây, phương pháp này cho thấy ổn định và hiệu quả tốt hơn.



Hình 1-2: Các phương pháp dựa trên phân tích cấu trúc màu

Các đặc trưng được trích xuất từ các phương pháp trên dựa trên các đặc trưng của kết cấu đều ở mức thấp, điều này chắc chắn sẽ ảnh hưởng đến độ mạnh mẽ và khả năng khái quát của mô hình. Vì các đặc trưng cấp thấp thường tồn tại trong không gian nhiều chiều và dễ bị nhiễu, chúng không có lợi cho việc phân loại trực tiếp [7]. Để cải thiện khả năng biểu diễn hình ảnh, hiệu quả nhận dạng và khả năng tổng quát của thuật toán, đồng thời hy vọng rằng các đặc trưng trong cùng một lớp sẽ giống nhau hơn, trong khi các đặc trưng giữa các lớp khác nhau thì khác biệt hơn. Vì vậy, mục tiêu sẽ là cần biểu diễn các đặc trưng cấp thấp tới các đặc trưng cấp cao khác biệt hơn thông qua một số thuật toán mã hóa. Các đặc trưng cấp cao có thể biểu diễn tốt hơn thông tin của toàn ảnh và giúp phân loại dễ dàng hơn.

Peixoto và cộng sự. [8] đầu tiên sử dụng bộ lọc DoG để thu được thông tin dải tần trung bình trong hình ảnh, sau đó trích xuất các đặc trưng chính thông qua biến đổi Fourier, và cuối cùng phân loại thông tin đặc trưng được trích xuất và xử lý thông qua bộ phân loại hồi quy logistic, để đạt được mục tiêu hình ảnh cần nhận dạng là khuôn mặt thật hay khuôn mặt giả mạo. Zhang và cộng sự. [9] đã đề xuất một sơ đồ phát hiện giả mạo khuôn mặt dựa trên đặc trưng Markov kết cấu màu (Colour Texture Markov feature - CTMF) và máy vector hỗ trợ loại bỏ tính năng đệ quy (SVM-RFE). Tác giả đã phân tích sự khác biệt giữa các pixel liền kề của khuôn mặt thật và khuôn mặt giả mạo, đồng thời xem xét đầy đủ thông

tin kết cấu giữa các kênh màu. Đầu tiên, sự khác biệt về kết cấu của khuôn mặt thật và mặt giả được ghi lại bằng bộ lọc vi phân có hướng (directional differential filter), có thể được coi là các đặc trưng cấp thấp của CTMF. Sau đó, tiến trình Markov được sử dụng để mô hình hóa sự khác biệt về kết cấu khuôn mặt để tạo thành biểu diễn cấp cao cho các đặc trưng cấp thấp. Cuối cùng, SVM-RFE được sử dụng để làm cho đạt được khả năng phát hiện trong thời gian thực.

Nhìn chung, phương pháp dựa trên phân tích kết cấu của hình ảnh có nhiều ưu điểm như chi phí thấp, thuật toán đơn giản và dễ thực hiện. Tuy nhiên, với sự phổ biến của máy ảnh độ nét cao và việc ứng dụng mặt nạ 3D chất lượng cao, việc sử dụng thông tin kết cấu không còn đáp ứng được nhu cầu nữa, do đó, thông tin kết cấu thường cần được tích hợp với các thông tin khác.

1.2.2. Phương pháp dựa trên tương tác người và máy

Con người có thể thực hiện các cử động hoặc tạo ra âm thanh theo yêu cầu, chẳng hạn như gật đầu, chớp mắt, mở miệng, mỉm cười, lè lưỡi, đọc một đoạn văn bản, trong khi sử dụng một khuôn mặt giả mạo thì những điều trên sẽ khó thực hiện. Dựa trên quan sát này, một phương pháp phát hiện giả mạo khuôn mặt người qua tương tác đã được đề xuất.

Phương pháp phát hiện giả mạo khuôn mặt qua tương tác ban đầu được thiết kế để cố định, cho phép việc video chuyển động được ghi sẵn có thể vượt qua loại thuật toán phát hiện giả mạo khuôn mặt này một cách dễ dàng. Để giải quyết vấn đề này, tính năng phát hiện tương tác giữa người và máy tính dựa trên các hướng dẫn chuyển động ngẫu nhiên ra đời. Tính ngẫu nhiên của hướng dẫn chuyển động khiến kẻ tấn công khó quay video trước để tấn công thuật toán phát hiện giả mạo khuôn mặt, điều này giúp cải thiện đáng kể hiệu suất phát hiện của thuật toán.

Wang và cộng sự. [10] đã tiến hành nhận dạng ngôn ngữ môi bằng cách phát hiện phạm vi thay đổi trong vùng miệng của khuôn mặt, được bổ sung bằng nhận dạng giọng nói để thu được thông tin giọng nói về phản ứng của người dùng để cùng đánh giá xem người dùng có đọc các câu được đưa ra ngẫu nhiên theo yêu cầu hay không. Singh và cộng sự. [11] sử dụng cử động chớp mắt và miệng để đưa ra các phán đoán sự sống. Diện tích của mắt cùng màu sắc, độ bão hòa, giá trị của răng được tính toán để xác định xem mắt có

mở và răng có hở hay không. Các đối tượng hành động theo cụm từ gợi ý do hệ thống tạo ngẫu nhiên và hoàn thành các hành động để chứng minh rằng đó là khuôn mặt thật. Ng và công sự [12] đã thiết kế một hệ thống tương tác máy tính với con người để hướng dẫn người dùng hoàn thành các biểu cảm ngẫu nhiên trên khuôn mặt. Bằng cách tính toán SIFT của nhiều khung hình ảnh, người dùng có thể được đánh giá liệu các biểu cảm khuôn mặt được chỉ định đã hoàn thành hay chưa và liệu chúng có phải là khuôn mặt thật hay không.

Phương pháp dựa trên tương tác giữa người và máy tính có thể làm giảm sự ảnh hưởng khi thay đổi các phương thức giả mạo một cách hiệu quả, hay nói cách khác phương pháp này khá tổng quát qua việc thực hiện thuật toán thông qua các hành động tương tác được thiết kế cẩn thận. Do đó, nó có tỷ lệ nhận dạng cao và tính linh hoạt tốt. Hiện nay, nó được sử dụng rộng rãi trong các tình huống kinh doanh thực tế như an ninh công cộng, điều trị y tế và tài chính. Tuy nhiên, phương pháp phát hiện giả mạo khuôn mặt dựa trên tương tác giữa con người và máy tính cần phải nhận biết liệu người dùng có hoàn thành hành động từ nhiều hình ảnh liên tiếp hay không và yêu cầu khả năng tính toán lớn và thời gian đưa ra kết luận lâu hơn so với thuật toán dựa trên chỉ một hình ảnh đơn lẻ. Hơn nữa, nó đòi hỏi sự hợp tác cao của chủ thể, quy trình phát hiện rườm rà và trải nghiệm người dùng chưa tốt nên vi phạm tính tiện lợi và lợi thế tự nhiên của công nghệ nhận dạng khuôn mặt.

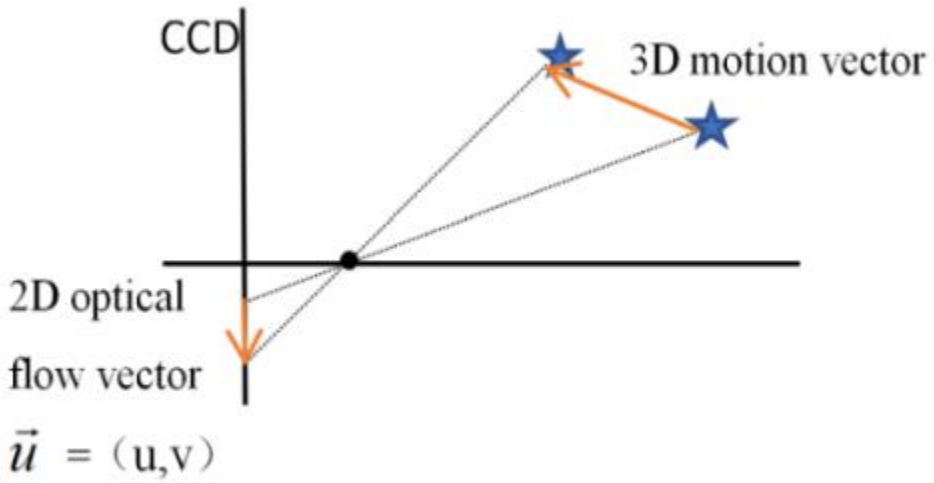
1.2.3. Các phương pháp dựa trên thông tin sự sống

Một sự khác biệt rõ ràng giữa khuôn mặt thật và khuôn mặt giả mạo là khuôn mặt thật có các đặc điểm quan trọng như nhịp tim, lưu lượng máu và chuyển động vi mô của cơ mặt không tự chủ và hầu hết các loại khuôn mặt giả rất khó bắt chước các đặc điểm sống đó. Phương pháp dựa trên thông tin sự sống chủ yếu sử dụng sự khác biệt trong các đặc điểm quan trọng này để phân biệt giữa khuôn mặt thật và khuôn mặt giả mạo.

1.2.3.1. Phân tích optical flow của khuôn mặt

Khái niệm optical flow được Gibson đề xuất lần đầu tiên vào năm 1950. Khi mắt người quan sát vật thể chuyển động, cảnh vật đó tạo thành một chuỗi hình ảnh thay đổi liên tục trên võng mạc của mắt người. Chuỗi thông tin thay đổi liên tục này liên tục "chảy" qua võng mạc (tức là mặt phẳng hình ảnh), giống như một loại "luồng" ánh sáng, nên được gọi

là optical flow. Khi một đối tượng đang chuyển động, hình dạng ánh sáng của điểm tương ứng trên ảnh cũng chuyển động, chúng ta có thể sử dụng optical flow để mô tả chuyển động của độ sáng ảnh, như trong hình 1-3. Có sự khác biệt trong các dạng chuyển động của khuôn mặt 3D và khuôn mặt 2D. Khi khuôn mặt quay và lắc, khuôn mặt thật tạo ra các luồng ánh sáng khác nhau do các chuyển động của khuôn mặt không nhất quán. Tuy nhiên, các chuyển động của khuôn mặt giả sử dụng ảnh về cơ bản là giống nhau, và luồng ánh sáng khá khác so với khuôn mặt thật. Dựa trên những khác biệt này, thông tin optical flow có thể được sử dụng để đưa ra phán đoán về mặt thật và mặt giả. Smiatacz và cộng sự. [13] đã tính toán các giá trị lưu lượng quang học được tạo ra bởi khuôn mặt khi quay, sau đó huấn luyện và phân loại các giá trị lưu lượng quang này bằng SVM. Bao et al. [14] đã sử dụng đường lưu lượng quang học (Optical Flow Line - OFL) để tính toán sự khác biệt không gian-thời gian của hình ảnh khuôn mặt người từ hai chiều ngang và dọc, đồng thời tổng hợp thông tin chuyển động của khuôn mặt người để phát hiện khuôn mặt giả mạo sử dụng ảnh và video. Phương pháp này tương đối đơn giản, nhưng nhạy cảm với ánh sáng và có hiệu quả phát hiện kém đối với giả mạo bằng video và giả mạo bằng mặt nạ 3D. Bởi vì phương pháp lưu lượng quang học tuân theo hai giả thiết: (1) Độ sáng không đổi; (2) Những chuyển động nhỏ khó đáp ứng được trong kịch bản đời thực, vì vậy chúng cũng có tác động nhất định đến hiệu quả phát hiện.



Hình 1-3: Phương pháp optical flow

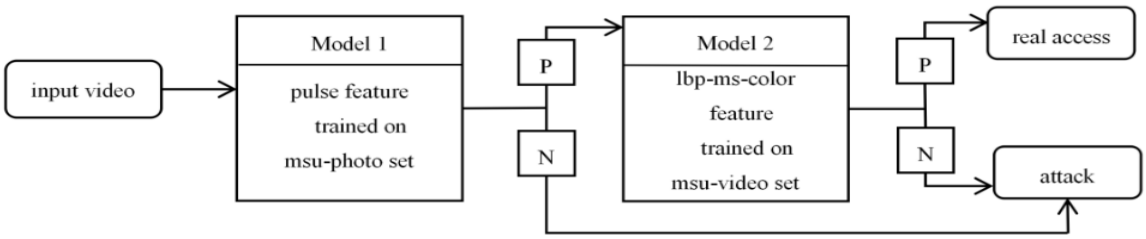
1.2.3.2. Phân tích phát hiện nhịp tim

Chụp cắt lớp nhân tạo (PPG) là một phương pháp phát hiện nhịp tim của cơ thể người bằng kỹ thuật chụp cắt lớp nhân tạo. Cách kiểm tra thường cần tiếp xúc trực tiếp. Đo nhịp tim không cần tiếp xúc sử dụng một máy ảnh, thường được gọi là chụp cắt lớp vi tính từ xa (Remote Photoplethysmography - rPPG). Khuôn mặt thật có nhiều mao mạch, và nhịp đập của tim sẽ dẫn đến những thay đổi về lưu lượng máu và vận tốc trong mạch máu, trong khi những thay đổi về lưu lượng máu sẽ ảnh hưởng đến sự hấp thụ và phản xạ ánh sáng trên khuôn mặt. Cuối cùng, những thay đổi như vậy trong máu sẽ dẫn đến sự thay đổi màu sắc của khuôn mặt. Sự thay đổi nhịp tim có thể thu được bằng cách trích xuất sự thay đổi màu sắc của vùng mao mạch dồi dào trên khuôn mặt. Khuôn mặt thật và khuôn mặt giả có sự phân bố nhịp tim khác nhau trong miền tần số, bằng cách sử dụng điều này, chúng ta có thể phân biệt được khuôn mặt là thật hay giả.

Li và cộng sự. [2] là những người đầu tiên áp dụng rPPG để phát hiện in vivo. Đầu vào là các khung hình video, bước đầu tiên là trích xuất các đặc trưng của nhịp tim. Nếu kết quả phân biệt là cơ thể sống, thì các đặc trưng về kết cấu và màu sắc dùng LBP sẽ được trích xuất thêm để phân biệt cơ thể sống hay giả mạo bằng màn hình điện tử. Bởi sự phân bố nhịp tim của khuôn mặt người trong video trên màn hình tương tự như nhịp tim của cơ thể sống. Quy trình cụ thể được trình bày trong hình 1-4. Liu et al. [3] tin rằng mặc dù các phương pháp dựa trên rPPG hiện có đã đạt được kết quả tốt trong các bộ dữ liệu khác nhau, nhưng chúng có thể không đủ mạnh khi các tín hiệu rPPG bị nhiễu. Do đó, họ đã đề xuất một tính năng mới - tính năng tương ứng rPPG (rPPG correspondence feature - CFrPPG) để xác định chính xác nhịp tim từ các tín hiệu rPPG nhiễu. Để khắc phục hiện tượng nhiễu toàn cục, một chiến lược học là đưa âm thanh nhiễu toàn cục vào đặc trưng CFrPPG được đề xuất. Tính năng được đề xuất không chỉ vượt trội so với phương pháp giả mạo mặt nạ 3D dựa trên rPPG, mà còn có thể xử lý với với điều kiện thực tế có ánh sáng yếu và chuyển động của máy ảnh.

Loại phương pháp khai thác nhịp tim này chủ yếu được sử dụng để phát hiện mặt nạ 3D giả mạo khuôn mặt con người. Trong điều kiện được chiếu sáng liên tục, đối tượng

được kiểm tra bằng cách giữ nguyên tư thế và biểu cảm, phương pháp này có độ chính xác cao. Tuy nhiên, quá trình xử lý họ yêu cầu video HD đủ dài để trích xuất tín hiệu rPPG đủ tốt và tín hiệu rPPG dễ bị ảnh hưởng bởi ánh sáng xung quanh và chuyển động của đối tượng được kiểm tra. Phương pháp này có tính ổn định trung bình, do đó, thường cần kết hợp các đặc trưng và bộ phân loại khác để phát hiện đặc trưng của khuôn mặt giả mạo.



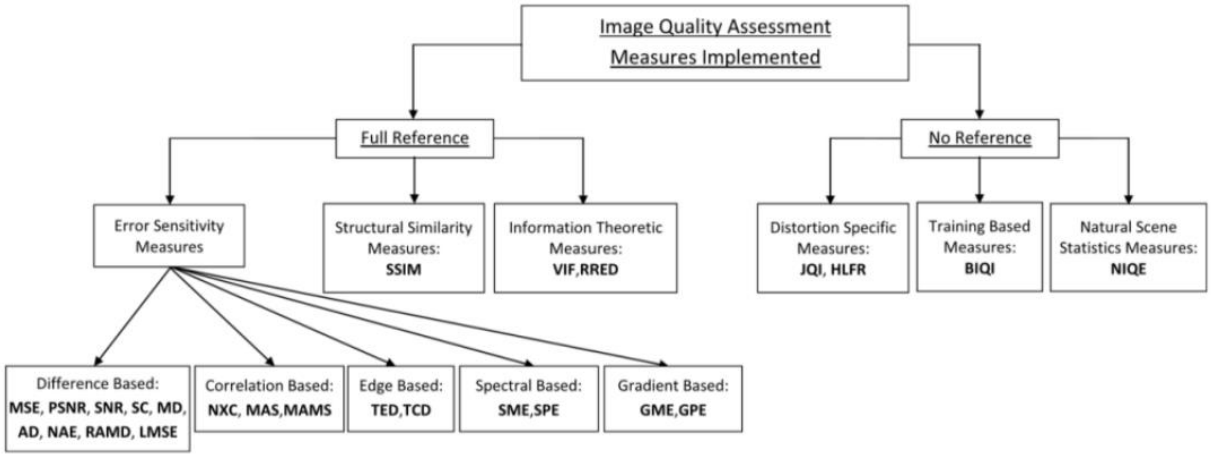
Hình 1-4: Cấu trúc của chiến lược cascadin

1.2.4. Các phương pháp dựa trên chất lượng của ảnh

Việc tạo một khuôn mặt giả mạo của một người cần một số thiết bị nhất định, cho dù đó là giấy ảnh, giấy in, thiết bị điện tử, silica gel, nhựa hoặc các phương tiện khác thì đặc tính chất liệu của chúng khác với các đặc điểm trên khuôn mặt và da của khuôn mặt sống. Sự khác biệt về chất liệu có thể dẫn đến sự khác biệt về tính chất phản chiếu, chẳng hạn như giấy ảnh, màn hình hiển thị điện thoại di động sẽ có một số phản xạ đặc trưng nhưng về cơ bản thì gương mặt thật sẽ không phản xạ gì. Mặc dù quy trình tạo ra một khuôn mặt giả mạo là tốt đến mấy thì hầu hết chất lượng hình ảnh sau khi chụp gián tiếp lần thứ 2 khác với khuôn mặt sống, chẳng hạn như sự phân bố màu sắc bị biến dạng và hình ảnh khuôn mặt giả mạo bị mờ đi. Các phương pháp dựa trên chất lượng hình ảnh chủ yếu sử dụng sự khác biệt giữa tính chất biến dạng và phản xạ của hình ảnh để phân biệt khuôn mặt thật và giả.

Galbally và cộng sự. [15] đề xuất đánh giá chất lượng hình ảnh bằng cách phân tích 25 yếu tố quan trọng trong các thước đo chất lượng hình ảnh (hình 1-5). Galbally và cộng sự. [16] cũng thiết kế 14 đặc trưng chung cho tính năng phát hiện giả mạo khuôn mặt để rút ra sự khác biệt về chất lượng hình ảnh. Lấy cảm hứng từ [15], Wen và cộng sự. [17] đã đề xuất một thuật toán phát hiện giả mạo khuôn mặt dựa trên phân tích biến dạng hình ảnh

(Image distortion analysis - IDA). Đầu tiên, bốn đặc trưng khác nhau (đặc điểm phản chiếu, mờ, khoảnh khắc màu và đa dạng màu) đã được trích xuất để tạo thành vectơ đặc trưng IDA. Sau đó, các bộ phân loại SVM được huấn luyện cho các cuộc tấn công giả mạo khuôn mặt khác nhau (chẳng hạn như giả mạo bằng ảnh và giả mạo video) tạo thành một bộ phân loại tích hợp để phân biệt khuôn mặt thật và giả. Cuối cùng, phương pháp này được áp dụng để phát hiện giả mạo khuôn mặt bằng nhiều khung hình dựa trên biểu quyết và thu được kết quả tốt. Chất lượng của hình ảnh phụ thuộc nhiều vào thiết bị chụp và điều kiện ngoại cảnh. Các điều kiện bên ngoài như thiết bị chụp chất lượng thấp và ánh sáng kém cũng có thể khiến hình ảnh khuôn mặt người thật bị biến dạng. Li và cộng sự. [18] đã xem xét ảnh hưởng của các thiết bị chụp có chất lượng khác nhau. Đầu tiên, các hình ảnh được phân nhóm theo kích thước chất lượng hình ảnh bằng phương pháp phân cụm, sau đó các mô hình phân loại hình ảnh dựa trên các đặc điểm chất lượng hình ảnh được huấn luyện cho từng mức chất lượng của hình ảnh. Đối với hình ảnh thử nghiệm, trước tiên xác định mức chất lượng hình ảnh của hình ảnh đó và sử dụng phương pháp hồi quy để tìm mô hình được huấn luyện cho chất lượng hình ảnh tương ứng. Sau đó dùng mô hình này để phân loại khuôn mặt thật hay giả mạo.



Hình 1-5: 25 yếu tố đánh giá chất lượng của hình ảnh

Phương pháp dựa trên chất lượng hình ảnh có độ phức tạp tính toán thấp và tốc độ phát hiện nhanh, có lợi thế trong việc phát hiện trong thời gian thực. Nhưng khi chất lượng hình ảnh cao, phương pháp này dễ bị vượt qua. Do đó, phương pháp cần khuôn mặt người thật với chất lượng cao hơn và hình ảnh khuôn mặt người giả mạo làm đầu vào để trích

xuất các đặc trưng chất lượng hình ảnh đủ tốt, đòi hỏi yêu cầu cao hơn đối với thiết bị thu thập hình ảnh của khuôn mặt.

1.2.5. Các phương pháp dựa trên thông tin về chiều sâu

Khuôn mặt thật có ba chiều, với thông tin độ sâu khác nhau ở các vị trí khác nhau như trán, mắt và đầu mũi, trong khi khuôn mặt ảnh và khuôn mặt video là hai chiều và thông tin độ sâu của các điểm khác nhau là như nhau. Ngay cả khi ảnh được gấp lại, nó có thông tin độ sâu khác với khuôn mặt thật, do đó, thông tin về chiều sâu có thể được sử dụng để chống giả mạo khuôn mặt.

Các phương pháp phát hiện giả mạo khuôn mặt dựa trên thông tin chiều sâu thường yêu cầu bổ sung các thiết bị phần cứng. Chất liệu để tạo một khuôn mặt giả mạo khác với vật chất ở da, mắt, môi và lông mày của khuôn mặt thật, và sự khác biệt này gây ra sự khác biệt về tính chất phản chiếu. Mặc dù khuôn mặt giả mạo trông rất giống với khuôn mặt thật trong điều kiện ánh sáng nhìn thấy, nhưng trong quang phổ hồng ngoại, sự xuất hiện của da, mắt, mũi và các vùng khác trên khuôn mặt thật khá khác so với khuôn mặt giả mạo. Một số nhà nghiên cứu đã sử dụng mô hình Gabor, HOG và lambert để trích xuất sự khác biệt phản chiếu giữa khuôn mặt thật và khuôn mặt giả mạo trong hình ảnh camera cận hồng ngoại để phát hiện giả mạo khuôn mặt [19–21]. Trong phổ hồng ngoại gần, khuôn mặt giả mạo trong ảnh và video khá khác với khuôn mặt thật, do đó, phương pháp này có độ chính xác cao, nhưng mặt nạ được chế tạo tinh xảo thì không khác mấy so với khuôn mặt sống. Để xác định các cuộc tấn công bằng mặt nạ, Steiner và cộng sự [22] đã sử dụng tia hồng ngoại sóng ngắn để phân biệt da mặt với mặt nạ. Ngoài ra, chúng ta cũng có thể sử dụng hình ảnh chiều sâu do camera chiều sâu (depth camera) chụp để ghi lại thông tin chiều sâu giữa các đối tượng để phát hiện giả mạo khuôn mặt. Wang và cộng sự. [23] kết hợp thông tin chiều sâu của máy ảnh Kinect và các đặc điểm cấu trúc học được từ mạng nơ-ron phức hợp để đánh giá khuôn mặt thật và giả, và cũng thu được kết quả tốt.

Nhìn chung, phương pháp phát hiện giả mạo khuôn mặt dựa trên thông tin độ sâu có những ưu điểm rõ ràng: thông tin độ sâu có đặc tính bất biến về độ chiếu sáng, do đó khả năng phát hiện giả mạo khuôn mặt là tốt; bản đồ độ sâu khuôn mặt thật có các đặc điểm

đường nét của khuôn mặt ba chiều, và có sự khác biệt đáng kể giữa bản đồ độ sâu của khuôn mặt chụp từ ảnh và khuôn mặt quay từ video; không có sự tương tác quá mức của người dùng, nó có tác dụng phát hiện tốt các cuộc tấn công bằng ảnh và video, nhưng việc phát hiện cuộc tấn công bằng mặt nạ 3D cần được nghiên cứu thêm. Tuy nhiên, các phương pháp này cần thêm phần cứng mới, đồng nghĩa với việc đầu tư thiết bị mới đắt tiền và cũng sẽ giới hạn phạm vi của thuật toán ở một mức độ nào đó.

Các phương pháp được đề cập ở trên đều dựa trên các đặc trưng nhân tạo. Mặc dù một số trong số chúng có thể đạt được tỷ lệ nhận dạng tốt hơn để phát hiện giả mạo khuôn mặt, nhưng vẫn còn một số nhược điểm, ví dụ như hiệu quả phát hiện phụ thuộc vào việc trích xuất các đặc trưng, nhu cầu đầu tư thêm phần cứng cũng như khả năng tổng quát và mạnh mẽ của thuật toán bị giới hạn.

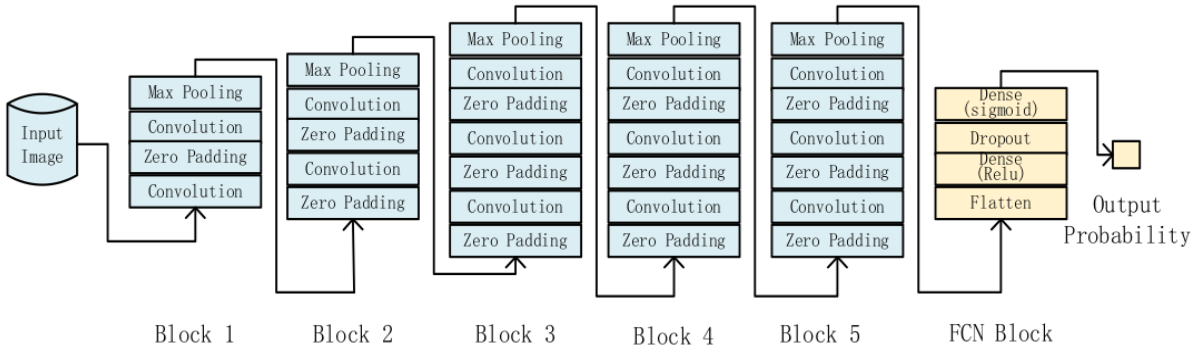
1.2.6. Các phương pháp dựa trên học sâu

1.2.6.1. Học chuyển tiếp (transfer learning)

Sử dụng học sâu để phát hiện khuôn mặt thật hay giả thường đòi hỏi một lượng lớn dữ liệu đào tạo để có được các đặc trưng khác biệt. Tuy nhiên, việc không có đủ dữ liệu cho bài toán phát hiện khuôn mặt và các mạng nơ-ron được dùng trong các phương pháp này chỉ bao gồm một vài lớp nên rất khó để huấn luyện một bộ phân loại mạng lớn với hiệu suất cao. Khi không có đủ dữ liệu để đào tạo từ đầu, học chuyển giao [24] có thể tránh được sự thích ứng quá mức (over-adapting) với các mạng có kiến trúc lớn và tiết kiệm nhiều tài nguyên máy tính.

Oeslle và cộng sự. [25] đã xây dựng một mô hình có tên FASNet và sử dụng mạng nơ-ron tích chập (CNN) được huấn luyện từ trước để phát hiện giả mạo khuôn mặt. Như thể hiện trong hình 1-6, FASNet đã sửa cấu trúc mạng trước đó trên cơ sở VGG16 [26] và sửa đổi ba lớp mạng cuối cùng để đạt được khả năng học chuyển tiếp. Đối với CNN, có hai phương pháp để học chuyển tiếp. Một cách đơn giản là sử dụng mô hình nguồn làm trình trích xuất đặc trưng sẵn có, sử dụng đầu ra của lớp đã chọn làm đầu vào cho mô hình đích, đây cũng là mô hình huấn luyện cho bài toán mới. Một cách tiếp cận phức tạp hơn là

đạt được sự "tinh chỉnh" (fine-tuning) toàn bộ hoặc một phần của mô hình nguồn bằng cách huấn luyện lại các trọng số thông qua lan truyền ngược.



Hình 1-6: Kiến trúc của FASNet

Tu và cộng sự. [27] đã đề xuất một mô hình data-driven hyper-depth đầy đủ dựa trên việc học chuyển giao. Mô hình này sử dụng mạng ResNet-50 [28] đã được huấn luyện từ trước để trích xuất các đặc trưng không gian của chuỗi các khung trình, sau đó truyền các đặc trưng không gian vào mạng nơ-ron LSTM để có được các đặc trưng thời gian, những đặc trưng này có thể được dùng cho kết quả phân loại cuối cùng để đánh giá các mặt thật hay giả.

Sử dụng học chuyển giao để huấn luyện khuôn mặt có thể giải quyết vấn đề thích ứng quá mức của mạng có kiến trúc lớn do hạn chế về tập dữ liệu. Khi trích xuất các đặc trưng chính có thể phân biệt khuôn mặt giả mạo với khuôn mặt thật, phương pháp này có thể giảm overfitting, thu được hiệu quả phát hiện tuyệt vời và tiết kiệm chi phí tính toán. Hiện các nghiên cứu về học chuyển giao trong bài toán phát hiện giả mạo là chưa nhiều và hiệu quả chưa đạt được kết quả mong đợi, lý tưởng nhất.

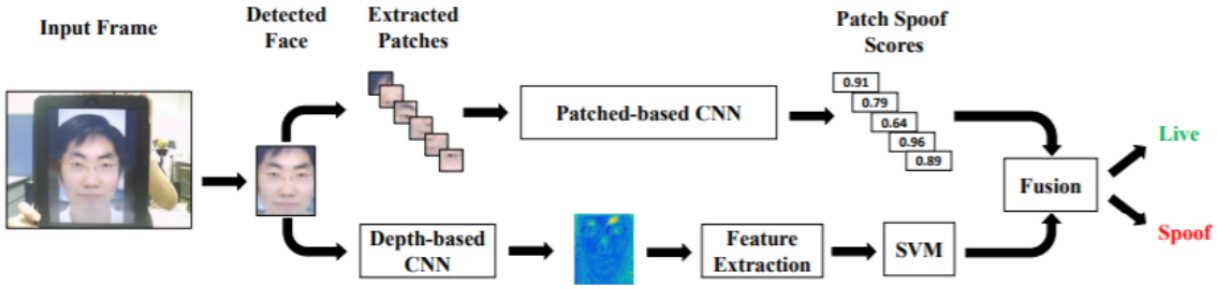
1.2.6.2. Kết hợp các đặc trưng

Một đặc trưng thì chỉ nhạy cảm với sự thay đổi của đặc trưng tương ứng của hình ảnh. Khi đặc trưng giữa khuôn mặt thật và khuôn mặt giả mạo ít có sự khác biệt thì thuật toán phân loại sẽ khó phân biệt chúng dựa trên việc huấn luyện đặc trưng đơn lẻ. Bằng cách trích xuất nhiều đặc trưng từ hình ảnh khuôn mặt, sự khác biệt giữa khuôn mặt thật

và khuôn mặt giả mạo có thể được làm nổi bật hơn, độ mạnh và khả năng tổng quát của thuật toán có thể được cải thiện và độ chính xác có thể được cải thiện đáng kể.

1.2.6.2.1. Thông tin về kiến trúc và chiều sâu

Thông tin về chiều sâu của hình ảnh là cơ sở quan trọng để đánh giá tính xác thực của khuôn mặt. Bởi vì khuôn mặt thật là ba chiều, trong khi khuôn mặt giả mạo bởi ảnh và màn hình là phẳng. Ngay cả khi khuôn mặt bị bóp méo, độ sâu trên khuôn mặt giả vẫn khác với khuôn mặt thật. Atoum và cộng sự. [29] lần đầu tiên lấy bản đồ chiều sâu khuôn mặt làm thông tin chính để phân biệt giả mạo khuôn mặt. Một phương pháp chống giả mạo khuôn mặt dựa trên hai kênh CNN đã được đề xuất để tích hợp các đặc trưng cục bộ của hình ảnh khuôn mặt với thông tin về chiều sâu. Mạng CNN đầu tiên trích xuất một số vùng khuôn mặt cục bộ làm dữ liệu huấn luyện, trả lại một điểm số cho mỗi vùng để thể hiện khả năng khuôn mặt đó là thật và tính toán toàn bộ hình ảnh khuôn mặt với giá trị trung bình. CNN thứ hai áp dụng mạng nơ-ron tích chập để ước tính bản đồ chiều sâu của hình ảnh khuôn mặt bằng cách phân loại các điểm pixel và cung cấp điểm số ước lượng độ thật của khuôn mặt theo bản đồ chiều sâu. Cuối cùng, điểm số của hai mạng CNNS này được tích hợp để đánh giá tính thật của khuôn mặt. Luồng xử lý của thuật toán được thể hiện trong hình 1-7. Mặc dù cách tiếp cận này đã cố gắng để tích hợp các đặc trưng, nhưng nó vẫn chưa vượt trội so với các phương pháp truyền thống.



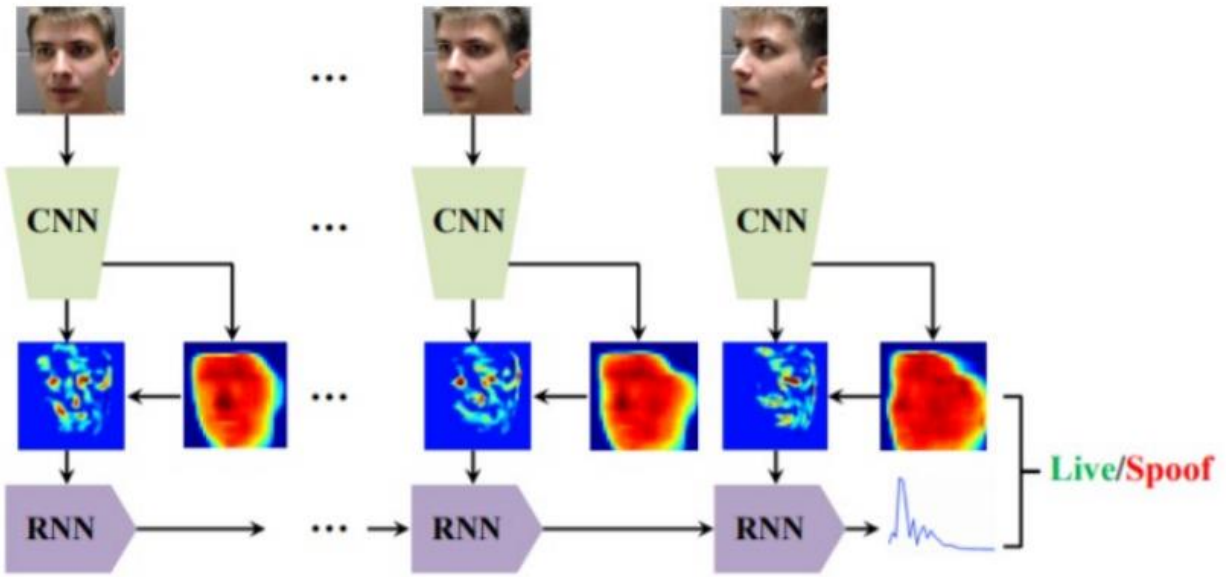
Hình 1-7: Biểu đồ luồng xử lý của CNN có 2 kênh

1.2.6.2.2. Thông tin không gian, thời gian

Hình ảnh khuôn mặt chứa một số lượng lớn các đặc trưng không gian như kết cấu và chiều sâu, nhưng các đặc trưng về thời gian cũng đóng một vai trò quan trọng trong việc

phát hiện giả mạo khuôn mặt. Việc phân tích khuôn mặt người từ thời gian và không gian có thể tìm ra thông tin hiệu quả hơn và cải thiện độ chính xác của việc phát hiện khuôn mặt giả mạo.

Liu và cộng sự. [30] đã tích hợp thông tin chiều sâu khuôn mặt và tín hiệu rPPG để thực hiện phát hiện giả mạo khuôn mặt và chỉ ra rằng bài toán phân loại nhị phân đã được thay thế bằng bài toán học giám sát đặc trưng được định trước. Chiều sâu của khuôn mặt đại diện cho thông tin không gian và tín hiệu rPPG đại diện cho thời gian, có thể làm nổi bật sự khác biệt chính giữa khuôn mặt thật và giả mạo. Từ góc độ không gian, khuôn mặt thực là ba chiều trong khi ảnh hoặc khuôn mặt trên màn hình điện tử là hai chiều; Từ góc độ thời gian, khuôn mặt thật có thể phát hiện tín hiệu rPPG bình thường nhưng khuôn mặt giả mạo thì không. Để đạt được hai loại giám sát, tác giả đã thiết kế một phương pháp học sâu dựa trên cấu trúc CNN - RNN. CNN giám sát độ sâu của hình ảnh để xác định các đặc trưng kết cấu rất khó phát hiện, sau đó đưa chiều sâu được ước tính và bản đồ đặc trưng vào một lớp đăng ký không chặt mới (new non-rigid registration layer) và tạo một bản đồ đặc trưng mới, trong khi RNN sử dụng bản đồ đặc trưng mới đã tạo trước đó và rPPG cho huấn luyện. Cuối cùng, thông tin chiều sâu và thống kê nhịp tim thu được bằng cách theo dõi chuỗi các tín hiệu rPPG được hợp nhất. Dựa trên cơ sở này, người ta đã phân biệt được mặt thật và mặt giả. Kiến trúc được thể hiện trong hình 1-8. Thử nghiệm cho thấy rằng phương pháp này đã đạt được kết quả thử nghiệm lý tưởng, cuối cùng nó đã vượt qua phương pháp thử nghiệm truyền thống và nó cũng phản ánh tầm quan trọng của việc giám sát phụ trợ.

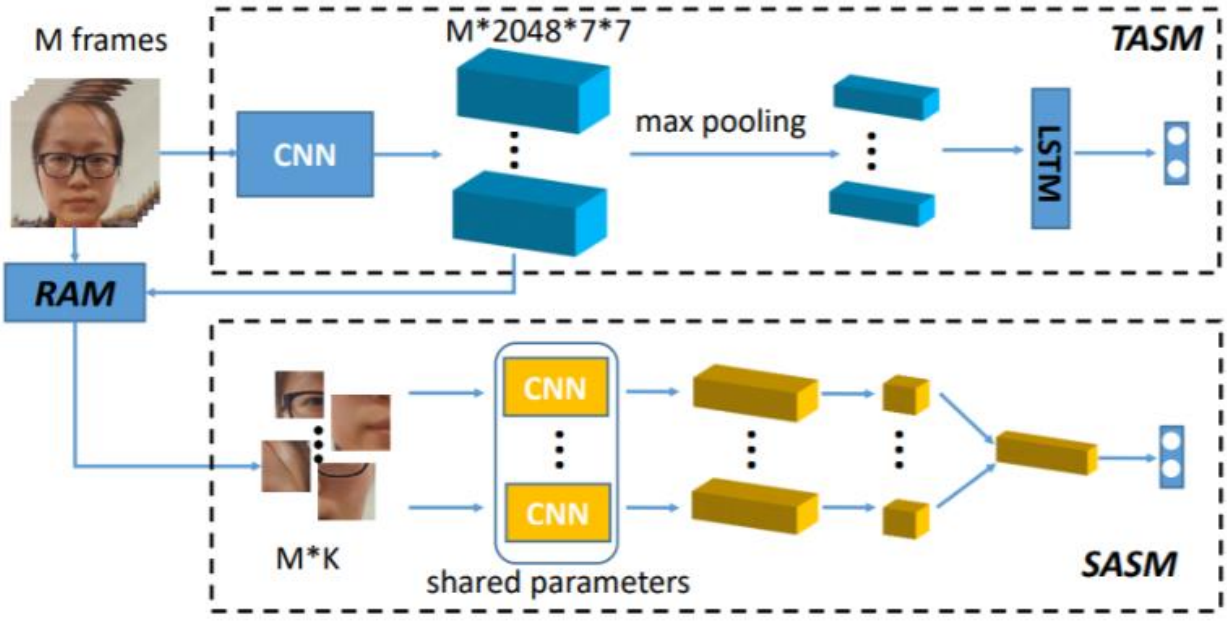


Hình 1-8: Phát hiện giả mạo khuôn mặt dựa trên CNN - RNN

Tuy nhiên, phương pháp trên cũng có những vấn đề sau: (1) Trong lớp đăng ký không chặt, các biểu hiện trên khuôn mặt và sự thay đổi tư thế bị loại bỏ, bỏ qua sự khác biệt của chúng giữa khuôn mặt thật và giả; (2) Việc sử dụng một khung hình để dự đoán chiều sâu là chưa thuyết phục. Một cách tương đối, việc tái tạo bản đồ chiều sâu thông qua những thay đổi vi mô không gian giữa nhiều khung hình ảnh có lợi hơn. Dựa trên hai thiếu sót này, Wang et al. [31] đã xây dựng một framework giám sát chiều sâu và sử dụng nhiều khung hình RGB như là đầu vào để ước lượng độ sâu của khuôn mặt để tận dụng đầy đủ thông tin không gian-thời gian, và phân tích ảnh hưởng của chuyển động và chiều sâu trong việc phát hiện các hình thức giả mạo khuôn mặt. Nó bao gồm hai mô-đun mới: Khối đặc trưng hướng dẫn luồng quang học (Optical Flow guided Feature Block - OFFB) và mô-đun ConvGRU, được thiết kế để trích xuất các chuyển động ngắn hạn và dài hạn để phân biệt các khuôn mặt thật với các khuôn mặt giả mạo. Phương pháp này có thể phát hiện khuôn mặt giả mạo một cách hiệu quả và chính xác.

Yang và cộng sự. [32] đã phát triển một mạng chống giả mạo dựa trên không gian-thời gian (STASN) mới, có tính đến thông tin không gian cục bộ và thời gian toàn cục để phân biệt khuôn mặt thật với khuôn mặt giả mạo. Mô hình bao gồm ba phần: TASM, RAM và SASM. TASM là CNN-LSTM, lấy chuỗi khung hình video làm đầu vào, đầu tiên trích

xuất các đặc trưng của CNN, thực hiện lan truyền LSTM và sau đó dự đoán kết quả. RAM học độ lệch dựa trên các đặc trưng CNN từ TASM và xuất ra các vùng tham gia liên quan đến chuỗi hình ảnh. SASM nhận đầu vào là vùng tham gia từ RAM và đầu ra đưa vào tham số chia sẻ CNN, và cuối cùng tích hợp nó để dự đoán, như thể hiện trong hình 1-9. Mô hình được đề xuất có thể tự động tập trung vào khu vực nhận dạng, giúp mô hình có thể phân tích hành vi. Bằng cách trích xuất các đặc trưng từ các khu vực khác nhau để tìm các dấu hiệu khó phát hiện, chẳng hạn như các cạnh, mẫu moire, v.v., mô hình có thể phân biệt hiệu quả khuôn mặt thật và giả mạo. Đồng thời, các tác giả nói rằng để chống giả mạo khuôn mặt, không chỉ để xây dựng một kiến trúc network tốt, dữ liệu cũng rất quan trọng. Do đó, họ đề xuất một giải pháp thu thập dữ liệu và công nghệ tổng hợp dữ liệu để mô phỏng các cuộc tấn công giả mạo khuôn mặt dựa trên phương tiện kỹ thuật số, có thể giúp thu được một lượng lớn dữ liệu huấn luyện phản ảnh được hình ảnh thực tế.



Hình 1-9: Kiến trúc của STASN

1.3. Kết luận

Các bài báo trên cho thấy việc tích hợp hai hay nhiều đặc trưng, đặc biệt là đặc trưng không gian - thời gian, làm cơ sở để đánh giá một khuôn mặt là thật hay giả có thể làm rõ sự khác biệt giữa hai khuôn mặt một cách toàn diện và hiệu quả hơn. So với đặc trưng phát hiện giả mạo khuôn mặt dựa trên một đặc trưng riêng lẻ, kết hợp nhiều đặc trưng có độ

chính xác vượt trội hơn, đồng thời cũng cải thiện tính mạnh mẽ và tổng quát của thuật toán. Tuy nhiên việc sử dụng nhiều khung hình cũng tăng thời gian và tài nguyên tính toán. Bên cạnh đó, các phương pháp sử dụng học sâu đang cho thấy điểm nổi trội hơn so với các phương pháp truyền thống.

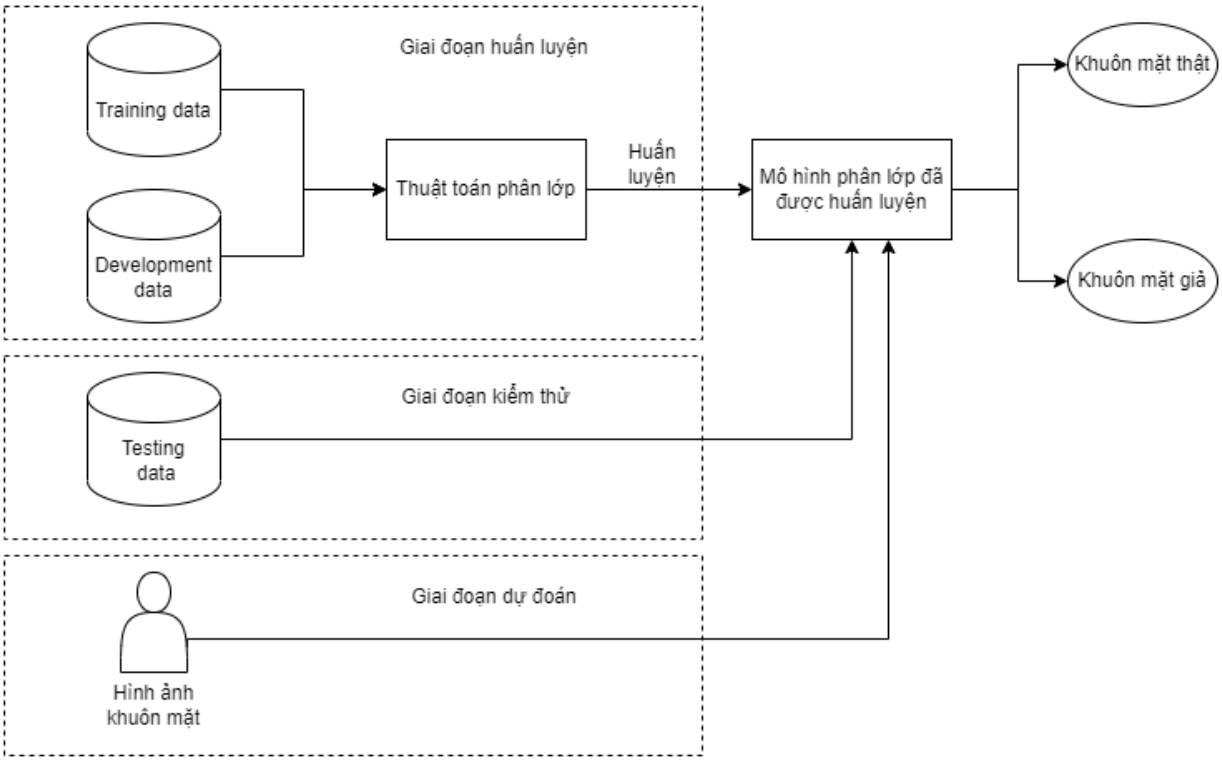
Như vậy, chương này đã giới thiệu về bài toán phát hiện giả mạo khuôn mặt và các nghiên cứu liên quan, cũng như cho thấy được tiềm năng của học sâu trong bài toán này. Chương tiếp theo sẽ trình bày về một số lý thuyết mạng học sâu và ý tưởng ứng dụng kỹ thuật này vào phát hiện giả mạo khuôn mặt.

CHƯƠNG 2: ỨNG DỤNG MẠNG HỌC SÂU VÀO BÀI TOÁN PHÁT HIỆN GIẢ MẠO KHUÔN MẶT

Chương này sẽ trình bày một số lý thuyết về mạng học sâu, ý tưởng đưa LBP vào mạng tích chập, sử dụng PRNet để tái tạo ảnh chiều sâu của khuôn mặt, giới thiệu về mạng resnet và cách kết hợp các kỹ thuật này lại với nhau để đưa ra được các phương pháp, kiến trúc mạng học sâu tương ứng cho bài toán phát hiện giả mạo. Đồng thời giới thiệu về vấn đề thích ứng miền và ý tưởng thử nghiệm nhằm khắc phục vấn đề này.

2.1. Ý tưởng giải quyết bài toán

Bài toán được đặt ra ban đầu trong luận văn đó là với một ảnh đầu vào, hệ thống phát hiện giả mạo cần trả về thông tin kết luận khuôn mặt xuất hiện trong bức ảnh đó là thật hay giả. Đây thực chất là một bài toán phân loại hai lớp. Vì vậy hướng tiếp cận của luận văn sẽ có 3 giai đoạn.



Hình 2-1: Các giai đoạn trong quá trình xây dựng giải pháp phát hiện giả mạo khuôn mặt

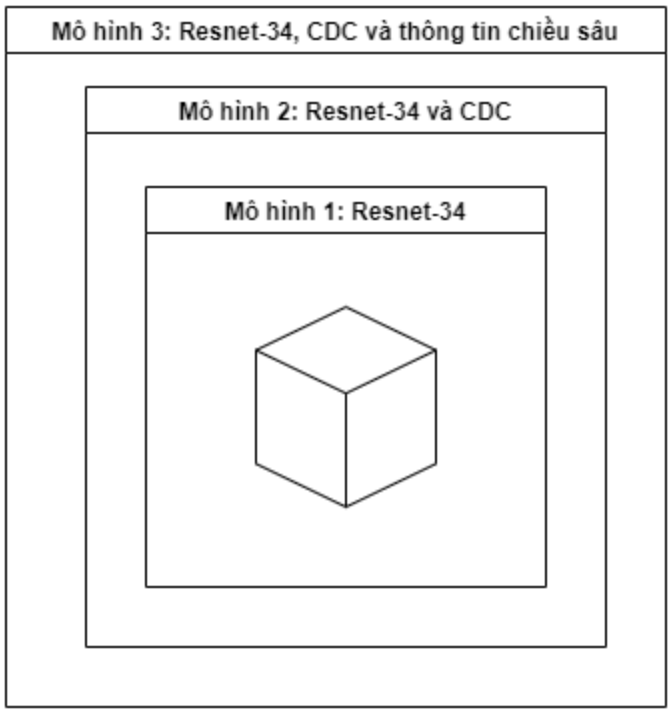
Giai đoạn huấn luyện: Đây là giai đoạn luận văn sử dụng tập dữ liệu huấn luyện và tập dữ liệu phát triển để huấn luyện mô hình phân lớp được lựa chọn.

Giai đoạn kiểm thử: Sau khi đã có được mô hình sau giai đoạn huấn luyện, luận văn sử dụng một tập dữ liệu kiểm thử chưa được sử dụng trong quá trình huấn luyện để đánh giá tổng quát mô hình vừa mới được tạo.

Giai đoạn dự đoán: Đây là giai đoạn tương lai, khi ứng dụng mô hình mà luận văn đã thử nghiệm vào trong các hệ thống đã được thực tế, đầu vào sẽ là một bức ảnh, đầu ra sẽ là dự đoán bức ảnh mới cho có phải là thật hay không.

Cụ thể hơn, trong phạm vi của luận văn, thuật toán phân lớp được chọn để thử nghiệm ở đây sẽ là một thuật toán học sâu bởi những thành tựu mà phương pháp này đạt được trong những năm gần đây.

Lấy cảm hứng từ các nghiên cứu được tìm hiểu ở chương 1, luận văn thấy rằng, có hai thông tin được sử dụng khá phổ biến để có thể phát hiện được khuôn mặt là thật hay giả mạo. Đó là đặc trưng LBP và thông tin về chiều sâu. Vì vậy, luận văn đã tìm hiểu các cách để đưa được các thông tin này vào một mạng học sâu là resnet-34. Cụ thể, các mô hình mà luận văn thử nghiệm sẽ được mở rộng dần dần với mô hình đầu tiên là sử dụng resnet-34 đơn lẻ để phân lớp, ở mô hình thứ 2 sẽ là mô hình 1 kết hợp với Central Difference Convolution (CDC), đây là một kỹ thuật kết hợp ý tưởng của LBP và phép toán tích chập. Cuối cùng, mô hình thứ 3 tạo bởi mô hình 2 và thông tin chiều sâu của khuôn mặt. Ở các mục tiếp theo luận văn sẽ mô tả kỹ hơn các kỹ thuật này cùng các kiến thức cơ bản về mạng học sâu.

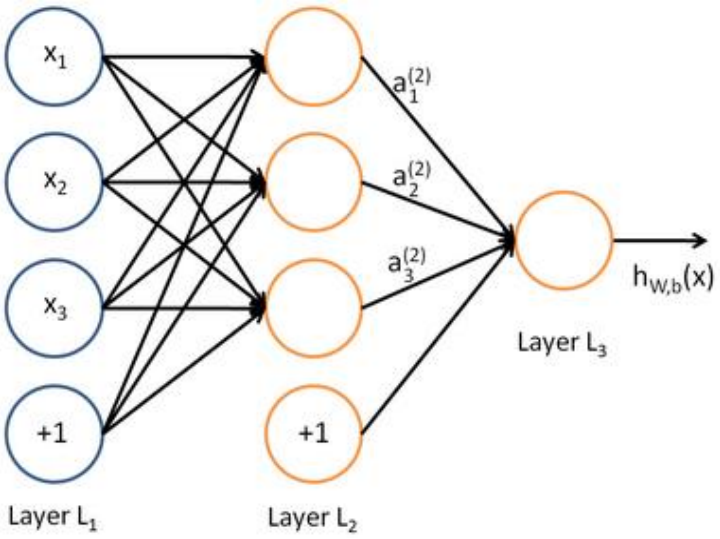


Hình 2-2: Các mô hình thử nghiệm làm bộ phân loại

2.2. Giới thiệu mạng học sâu

2.2.1. Mạng nơ-ron

Mạng nơ-ron (neural network) là một mô hình học máy có khả năng mô phỏng bất kỳ hàm số hay quan hệ nào [17]. Mạng nơ-ron có nhiều kiến trúc nhưng trong phạm vi luận văn này, nội dung sẽ giới thiệu về mạng nơ-ron có kiến trúc kết nối đầy đủ (fully connected network). Ở kiến trúc này mạng nơ-ron gồm nhiều lớp (layer) được sắp xếp tuyến tính, mỗi lớp gồm gồm nơ-ron và các nơ-ron này không kết nối với nhau, các nơ-ron ở 2 lớp liên tiếp được kết nối với nhau thông qua các cạnh, mỗi cạnh đều có một trọng số, tạo thành một ma trận trọng số để kết nối giữa 2 lớp liên tiếp.



Hình 2-3: Mạng Nơ-ron với giá trị đầu ra

Lớp đầu tiên bên trái được gọi là lớp đầu vào (input layer), lớp ngoài cùng bên phải được gọi là lớp đầu ra (output layer), lớp ở giữa được gọi là lớp ẩn (hidden layer). Vòng tròn có nhãn +1 được gọi là đơn vị thiên vị (bias unit). Tập hợp thông số của mạng là $(W, b) = \left(W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}\right)$ ở đây $W^{(l)}$là ma trận trọng số liên kết giữa 2 lớp l và $l + 1$, cũng như $b^{(l)}$là đơn vị thiên vị của lớp $l + 1$. Ở trong ví dụ này $W^{(1)}$là ma trận có kích thước 3 x 3 và $W^{(2)}$ có kích thước 1 x 3.

Biểu thị $a_i^{(l)}$là giá trị kích hoạt (đầu ra có đơn vị i trong lớp l). Với l = 1 có $a_i^{(l)} = x_i$. Cuối cùng, giả sử tập thông số của mạng (W, b) đã được xác định hết thì có đầu ra của mạng là $h_{W,b}$ là một số thực và được tính toán cụ thể như sau:

$$a_1^{(2)} = f\left(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)}\right)$$

$$a_2^{(2)} = f\left(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_2^{(1)}\right)$$

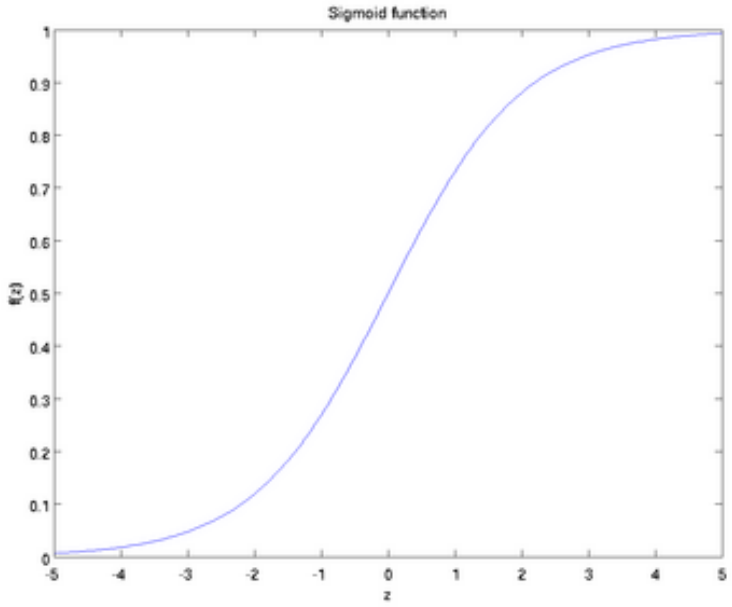
$$a_3^{(2)} = f\left(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_3^{(1)}\right)$$

$$h_{W,b}(x) = a_1^{(3)} = f\left(W_{11}^{(2)}a_1^{(2)} + W_{12}^{(2)}a_2^{(2)} + W_{13}^{(2)}a_3^{(2)} + b_1^{(2)}\right)$$

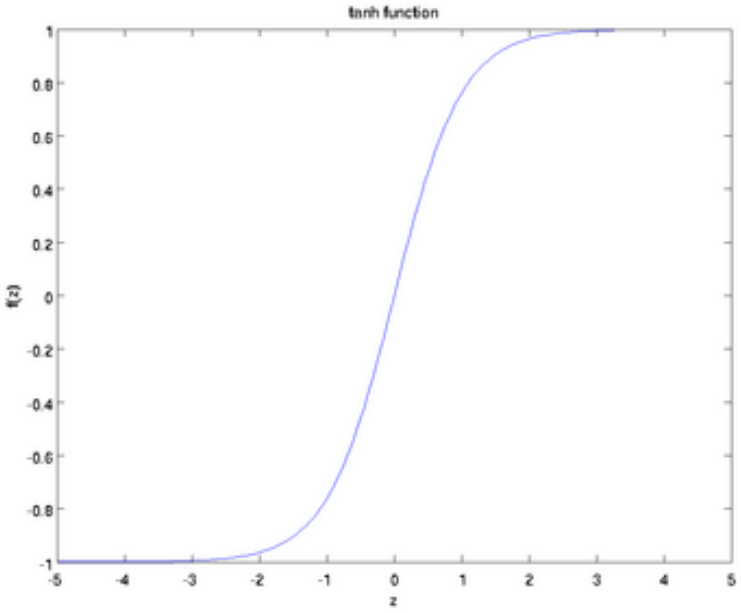
Chúng ta cũng có thể định nghĩa $z_i^{(l)} = \sum_{j=1}^{n} W_{ij}^{(l-1)}x_j + b_i^{(l-1)}$ vậy $a_i^{(l)} = f(z_i^{(l)})$

Trong đó hàm f được gọi là hàm kích hoạt (activation function). Một số lựa chọn cho hàm kích hoạt như:

- Hàm sigmoid: $f(z) = \frac{1}{1+exp\,(-z)}$
- Hàm tanh: $f(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$



Hình 2-4: Đồ thị của hàm sigmoid



Hình 2-5: Đồ thị của hàm tanh

Mạng nơ-ron được ứng dụng rộng rãi để giải quyết bài toán phân loại, do đó ở đầu ra của mạng cần phải là một hàm phân bố xác suất của các nhãn phân loại cho biết tỉ lệ phần trăm dữ liệu đầu vào có thể là nhãn đó. Giả sử có n nhãn để phân loại thì đầu ra ở lớp cuối có n giá trị z = $(z_1, z_2, \dots, z_n)$ và cần chuyển n giá trị này thành một phân bố xác suất bằng cách sử dụng hàm softmax như sau:

$$softmax(z) = (p_1, p_2, \dots, p_n)$$

Trong đó: $p_i = \frac{exp\ (z_i)}{\sum_{j=1}^{n} exp\ (z_j)}$ với $exp\ (.)$ là hàm lũy thừa cơ số tự nhiên e.

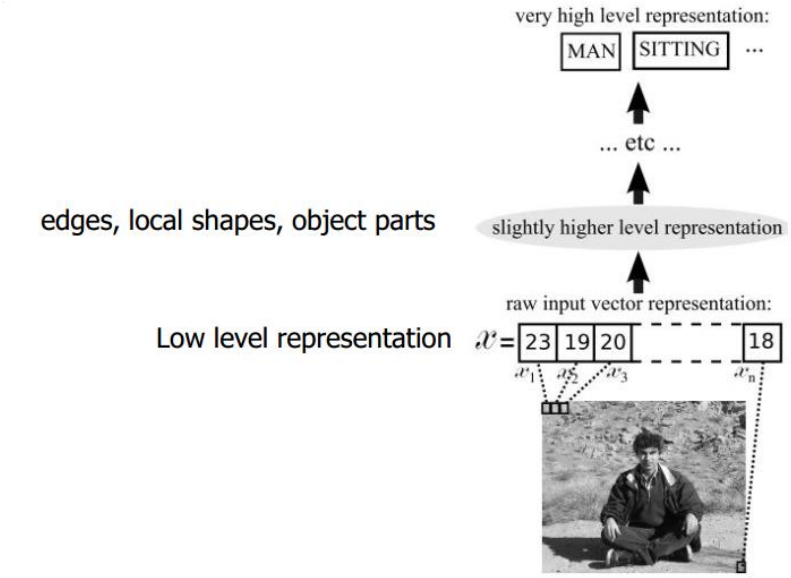
Như vậy, kết quả đầu ra của mạng nơ-ron phụ thuộc vào thông số (W, b), để đánh giá thông số (W, b) có tốt hay không, người ta có một giá trị gọi là hàm mất mát (Loss function) để tính độ sai lệch giữa dữ liệu ra của mạng và dữ liệu cần đạt được.

Dựa vào giá trị của hàm mất mát, quá trình huấn luyện mạng nơ-ron sẽ liên tục cập nhật các trọng số của mạng nơ-ron đó thông qua một thuật toán gọi là lan truyền ngược.

2.2.2. Học sâu

Học sâu là một nhánh của học máy dựa trên một tập các thuật toán cố gắng để có được các mô hình trừu tượng mức độ cao trong dữ liệu bằng cách sử dụng các đồ thị sâu với nhiều lớp xử lý bao gồm các biến đổi tuyến tính và phi tuyến tính.

Phương pháp này là một phần của các phương pháp học máy dựa trên việc học các đặc điểm, biểu hiện của dữ liệu. Một quan sát (ví dụ như một hình ảnh) có thể được biểu hiện bằng nhiều cách khác nhau như tập giá trị cường độ của các điểm ảnh, hoặc theo tập các cạnh, hay các vùng có hình dạng đặc biệt. Trong học sâu, các đặc trưng có đặc điểm có nhiều mức, các đặc trưng ở mức cao được học từ các đặc trưng ở mức thấp hơn.



Hình 2-6: Mô hình của một mạng học sâu

Nghiên cứu trong lĩnh vực này cố gắng để làm cho các đặc trưng, biểu hiện của dữ liệu tốt hơn và tạo ra các mô hình để học được các đặc trưng, biểu hiện này từ các dữ liệu không có nhãn ở quy mô lớn.

Nhiều kiến trúc học sâu khác nhau như mạng nơ-ron sâu, mạng nơ-ron tích chập sâu, mạng niềm tin sâu, mạng nơ-ron tái phát … đã được áp dụng cho các lĩnh vực như thị giác máy tính, tự động nhận dạng giọng nói, xử lý ngôn ngữ tự nhiên, nhận dạng âm thanh ngôn ngữ và tin sinh học, chúng đã được chứng minh là tạo ra các kết quả rất tốt đối với nhiều

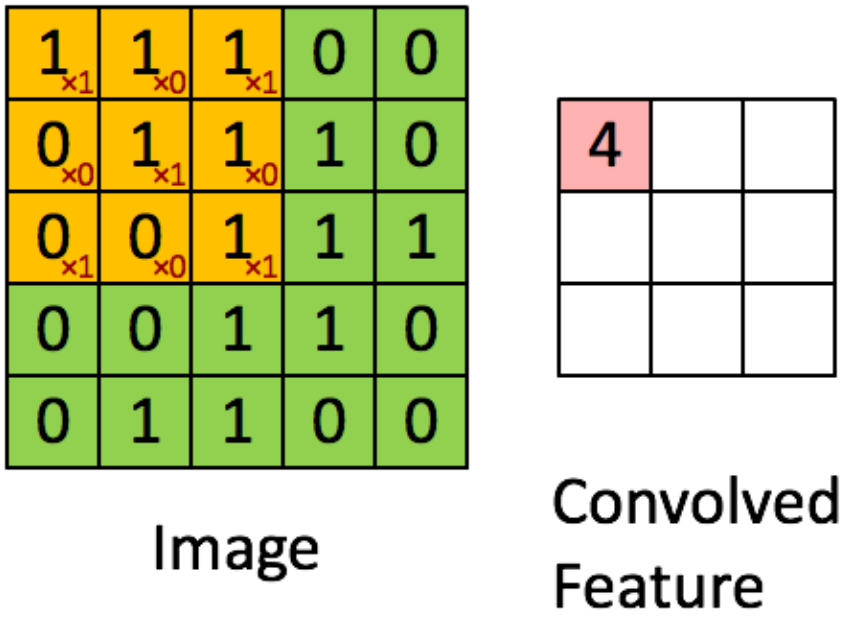
nhiệm vụ khác nhau. Tuy nhiên, mạng học sâu không phải là hoàn hảo hoàn toàn mà nó cũng có những nhược điểm nhất định:

- ***Tính sẵn có của dữ liệu*:** Thuật toán học chính của mạng học sâu là khởi tạo ngẫu nhiên trọng số của mạng, sau đó đào tạo nó bằng cách sử dụng một tập huấn luyện có nhãn, dùng một thuật toán học có giám sát ví dụ gradient descent để làm giảm lỗi huấn luyện. Đây cũng là một khó khăn của mạng học sâu bởi dữ liệu được gắn nhãn là khan hiếm và do đó rất khó để có đủ dữ liệu làm ví dụ để đào tạo được thông số cho một mô hình phức tạp như mạng học sâu. Việc dữ liệu không đủ có thể dẫn tới overfitting.
- ***Cực trị địa phương (Local optima)***: Huấn luyện một mạng nông (chỉ với 1 lớp ẩn) sử dụng học có giám sát thường dẫn đến kết quả các thông số hội tụ hợp lý. Nhưng khi huấn luyện một mạng học sâu, việc này ít nhiều cũng không tốt bằng. Đặc biệt, việc huấn luyện một mạng nơ-ron sử dụng học có giám sát có liên quan đến giải quyết bài toán tối ưu hóa hàm không lồi (cực tiểu hóa hàm $\sum_i \quad \|h_W(x^{(i)}) - y^{(i)}\|^2$ với thông số là W, b). Ở mạng học sâu, có rất nhiều cực trị địa phương xấu.
- ***Sực khếch tán của gradients (Diffusion of gradient):*** Có thêm một lý do mà giải thích tại sao gradient descent không hoạt động tốt trên một mạng học sâu với các trọng số được khởi tạo ngẫu nhiên. Cụ thể khi sử dụng thuật toán lan truyền ngược để tính đạo hàm, gradient được lan truyền ngược lại (từ lớp đầu ra về các lớp trước đó của mạng) sẽ nhanh chóng giảm bớt độ lớn khi độ sâu của mạng tăng lên. Kết quả là, đạo hàm của toàn bộ chi phí (overall cost) với trọng số của các lớp đầu tiên rất nhỏ. Bởi vậy khi sử dụng gradient descent, trọng số của các lớp này thay đổi chậm, và các lớp này học không được nhiều.

2.2.3. Mạng nơ-ron tích chập

2.2.3.1. Tích chập

Tích chập (convolution) hiểu một cách đơn giản nhất dùng một cửa sổ trượt để trượt trên một ma trận. Hình dưới đây sẽ cho một cách nhìn dễ hình dung hơn.



Hình 2-7: Ví dụ khi thực hiện tích chập trên ảnh

Tưởng tượng rằng ma trận bên trái thể hiện một bức ảnh màu đen và trắng. Mỗi ô tương ứng với một điểm ảnh, 0 cho màu đen và 1 cho màu trắng (hoặc là từ 0 đến 255 đối với ảnh đa mức xám). Cửa sổ trượt được gọi là hạt nhân (kernel), bộ lọc (filter) hoặc phát hiện đặc trưng (features detector). Ở đây sử dụng một bộ lọc kích thước 3 x 3, mỗi khi trượt bộ lọc trên ma trận, thực hiện nhân các giá trị tương ứng trên bộ lọc và giá trị trên ma trận đang trùng khớp nhau, sau cùng lấy tổng tất cả các giá trị đạt được. Làm tương tự như vậy khi trượt bộ lọc trên toàn bộ ma trận. Ví dụ:

Ở hình ảnh trên fillter đang được đặt trùng với ma trận con có tọa độ góc trên trái nằm ở ô (1, 1) của ảnh. Kết quả khi thực hiện như sau:

1 * 1 + 0 * 1 + 1 * 1 + 0 * 0 + 1 * 1 + 0 * 1 + 1 * 0 + 0 * 0 + 1 * 1 = 4

Dưới đây sẽ là một vài ví dụ cho việc sử dụng filter.

- Lấy trung bình giá trị điểm ảnh để làm mờ ảnh

s

0	0	0	0	0
0	1	1	1	0
0	1	1	1	0
0	1	1	1	0
0	0	0	0	0

Hình 2-8: Sử dụng bộ lọc để làm mờ ảnh

- Lấy sự khác biệt giữa điểm ảnh và các giá trị lân cận nó để phát hiện cạnh:

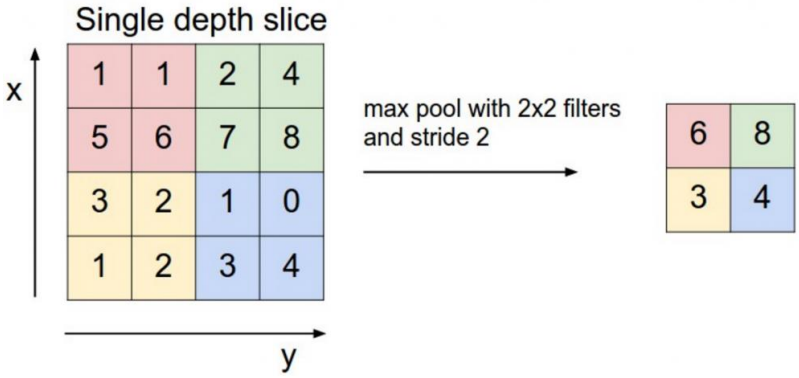
	0	1	0	
	1	-4	1	
	0	1	0	

Hình 2-9: Sử dụng bộ lọc để phát hiện cạnh

2.2.3.2. Mạng tích chập (Convolution neural networks - CNN)

Mạng tích chập là một mạng kết hợp giữa vài lớp tích chập (convolution layer) với các hàm kích hoạt phi tuyến như ReLu hoặc tanh áp dụng cho các kết quả có được sau convolution.

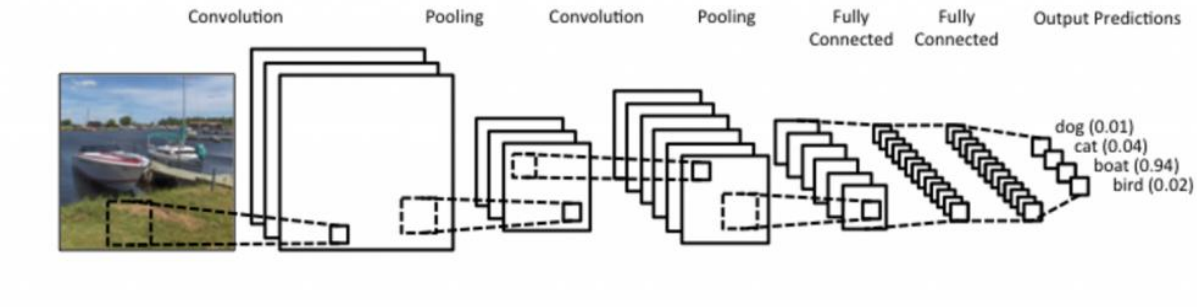
Trong mô hình mạng nơ-ron truyền thống (feedforward neural network), các lớp kết nối trực tiếp với nhau. Các lớp này còn được gọi là các kết nối đầy đủ (Fully connected layer). Trong mô hình CNN thì không như vậy, CNN sử dụng tích chập trên lớp đầu vào để tính toán đầu ra, mỗi lớp được dùng một bộ lọc khác nhau. Ngoài tích chập, CNN còn có các lớp pooling (pooling layer), lớp pooling cũng gần tương tự như tích chập để dễ hình dung, pooling cũng dùng một cửa sổ trượt, trượt trên một ma trận, nhưng thay vì nhân các phần tử đang khớp nhau trên bộ lọc và các phần tử trên ma trận và tính tổng các giá trị có được thì pooling có nhiều lựa chọn như lấy giá trị lớn nhất của các giá trị trên ma trận đang khớp với cửa sổ, hay lấy giá trị nhỏ nhất v…v nhưng thường sử dụng là lấy giá trị lớn nhất trên các phần tử của ma trận khi đó pooling có một tên gọi riêng là max pooling.



Hình 2-10: Ví dụ về thực hiện pooling

Pooling giúp giảm số chiều ở đầu ra nhưng hy vọng sẽ vẫn giữ được các thông tin nổi bật nhất.

Cuối các lớp tích chập, pooling là các lớp kết nối đầy đủ.



Hình 2-11: Mạng học sâu để phân loại đối tượng trong ảnh

Ở cuối cùng có một lớp softmax lớp softmax này có nhiệm vụ chuyển đổi các giá trị đầu vào thành một phân bố xác suất để dự đoán đối tượng, ví dụ như trong ảnh trên thì cuối cùng của mạng CNN này cho ra lần lượt xác suất của các nhãn đang cần dự đoán cho bức ảnh đầu vào, ví dụ như bức hình ở trên xác suất của đối tượng trong bức ảnh với các nhãn là:

- Dog: 0.01
- Cat: 0.04
- Boat: 0.94
- Bird: 0.02

Nhìn vào xác suất phía trên thấy được rằng đối tượng trong ảnh được dự đoán là thuyền.

2.3. Ứng dụng học sâu vào bài toán phát hiện giả mạo khuôn mặt

2.3.1. Mạng tích chập khác biệt trung tâm (Central Difference Convolution - CDC)

Phép toán tích chập 2D là một phép toán cơ bản của mạng tích chập trong các bài toán thị giác máy tính. Có hai bước trong quá trình thực hiện phép toán này. Bước đầu tiên là lấy mẫu, đây là bước thực hiện thu thập các giá trị vùng cục bộ R trên bản đồ đặc trưng x. Bước thứ 2 là tổng hợp các giá trị thu được từ bước 1 với trọng số. Cuối cùng, đầu ra là bản đồ đặc trưng y có thể được tính như sau:

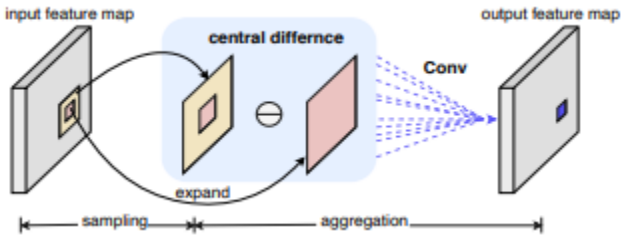
$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n),$$

Với p_0 là vị trí hiện tại ở các bản đồ đặc trưng đầu vào và đầu ra, p_n là tập các vị trí trong R. Ví dụ phép tích chập với kernel 3x3 có độ giãn là 1 thì có R = {(−1, −1), (−1, 0), · · ·, (0, 1), (1, 1)}.

Để có thể áp dụng tư tưởng của LBP, một phương pháp mô tả mối quan hệ địa phương vào phép toán tích chập nhằm mục đích tăng cường khả năng biểu diễn và khái quát hóa của mô hình. Việc đưa ý tưởng này vào phép toán tích chập để tạo thành CDC như sau: Tương tự mạng tích chập thông thường, CDC cũng bao gồm hai bước lấy mẫu và tổng hợp. Việc lấy mẫu tương tự như trong tích chập thông thường trong khi bước tổng hợp có khác biệt: như được minh họa trong hình 2-12, CDC lấy tổng center-oriented gradient của các giá trị được lấy mẫu. Phương trình (1) trở thành

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0)).$$

Khi p_n = (0, 0), giá trị gradient luôn bằng 0 đối với chính vị trí trung tâm p_0.



Hình 2-12: Mô hình của CDC

Đối với bài toán phát hiện giả mạo khuôn mặt, cả thông tin ngữ nghĩa mức cường độ và thông tin chi tiết mức gradient đều rất quan trọng để phân biệt khuôn mặt sống và khuôn mặt giả mạo, điều này chỉ ra rằng việc kết hợp tích chập thông thường với CDC có thể là một cách khả thi để cung cấp năng lực mô hình hóa mạnh mẽ hơn. Do đó CDC được tổng quát hóa thành:

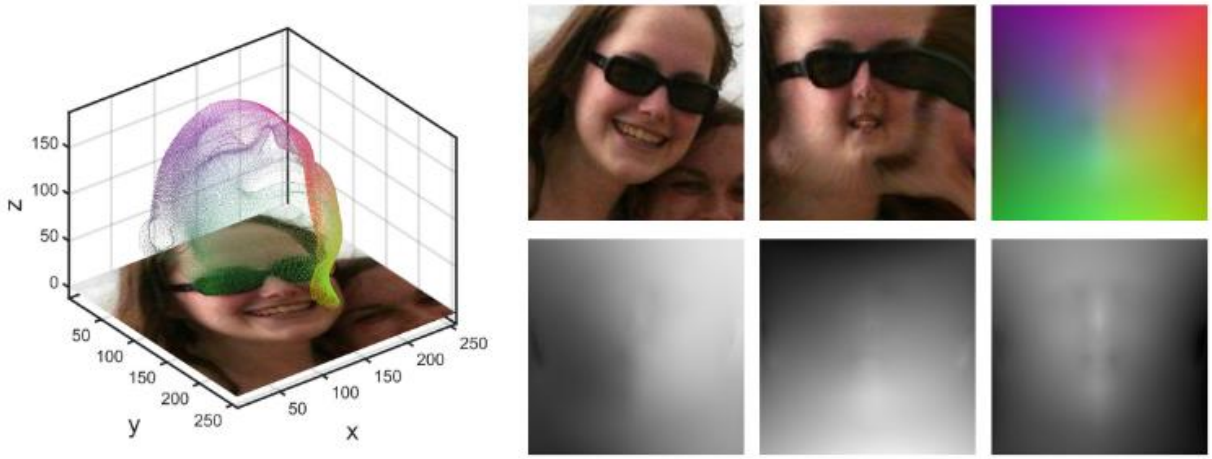
$$y(p_0) = \theta \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot (x(p_0 + p_n) - x(p_0))}_{\text{central difference convolution}} + (1-\theta) \cdot \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot x(p_0 + p_n)}_{\text{vanilla convolution}},$$

Với thông số θ (theta) thuộc đoạn [0, 1] thể hiện tỷ lệ đóng góp giữa thông tin mức cường độ và mức gradient. Các giá trị của θ càng cao có nghĩa là thông tin về độ chênh lệch trung tâm càng có tầm quan trọng.

2.3.2. Tạo thông tin chiều sâu từ khuôn mặt

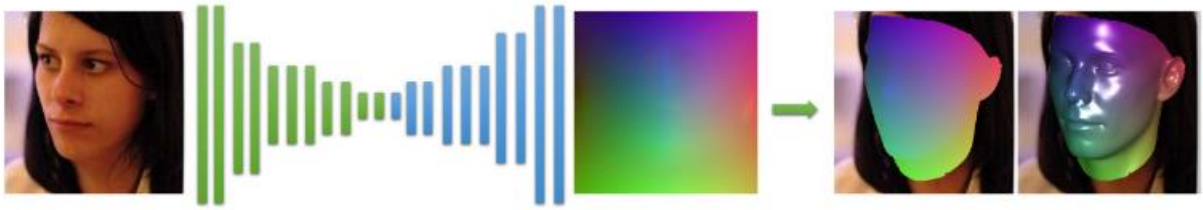
2.3.2.1. Biểu diễn khuôn mặt 3D

Để biểu diễn một khuôn mặt 3D trong máy tính, luận văn sử dụng một biểu đồ vị trí UV. Bản đồ vị trí UV hay gọi tắt là bản đồ vị trí, là một hình ảnh 2D ghi lại vị trí 3D của tất cả các điểm trong không gian UV. Trong những năm qua, không gian UV hoặc tọa độ UV, là một mặt phẳng hình ảnh 2D được tham số hóa từ không gian 3D, đã được sử dụng như một cách để thể hiện thông tin bao gồm kết cấu của khuôn mặt (bản đồ kết cấu) [35,36,37,38], hình học 2,5D [39,40] và sự tương ứng giữa các mắt lưới face 3D [41]. Trong luận văn, không gian UV được dùng để lưu trữ tọa độ 3D của các điểm từ mô hình khuôn mặt 3D. Như trong hình 2-13, chúng ta xác định đám mây điểm mặt 3D trong hệ tọa độ Descartes. Điểm gốc của không gian 3D chồng lên phía trên bên trái của hình ảnh đầu vào, với trục x dương hướng sang bên phải của hình ảnh. Đám mây điểm của khuôn mặt 3D khớp chính xác với khuôn mặt trong hình ảnh 2D khi được chiếu lên mặt phẳng x-y. Do đó, bản đồ vị trí có thể dễ dàng hiểu được là thay thế giá trị r, g, b trong bản đồ kết cấu bằng các tọa độ x, y, z.



Hình 2-13: Hình minh họa của bản đồ vị trí UV. Bên trái: Biểu đồ 3D của hình ảnh đầu vào và đám mây điểm 3D của khuôn mặt. Bên phải: Hàng đầu tiên là hình ảnh 2D đầu vào, bản đồ kết cấu UV được trích xuất và bản đồ vị trí UV tương ứng. Hàng thứ hai là kênh x, y, z của bản đồ vị trí UV.

2.3.2.2. Kiến trúc mô hình và hàm mất mát



Hình 2-14: Kiến trúc của PRNet. Các hình chữ nhật màu xanh lá cây đại diện cho các khối Resnet và các hình chữ nhật màu xanh đại diện cho các lớp tích chập đã chuyển vị.

Mô hình học sâu dùng để cấu trúc lại thông tin chiều sâu của khuôn mặt tên là PRNet. Vì mô hình PRN chuyển hình ảnh RGB đầu vào thành hình ảnh bản đồ vị trí nên mô hình sử dụng cấu trúc bộ mã hóa-giải mã để học hàm chuyển đổi. Phần bộ mã hóa trong PRN bắt đầu bằng một lớp tích chập, theo sau là 10 khối residual [42] giúp giảm hình ảnh đầu vào có kích thước 256 × 256 × 3 thành bản đồ đặc trưng có kích thước 8 × 8 × 512, phần bộ giải mã chứa 17 lớp tích chập chuyển vị để tạo ra dự đoán bản đồ vị trí có kích thước 256 × 256 × 3. Mô hình sử dụng kích thước hạt nhân (kernel size) là 4 cho tất cả các

lớp tích chập hoặc chuyển vị và sử dụng lớp ReLU để kích hoạt. Do bản đồ vị trí chứa cả thông tin 3D đầy đủ và kết quả căn chỉnh khuôn mặt. Kiến trúc của PRNet được thể hiện trong hình 2-14.

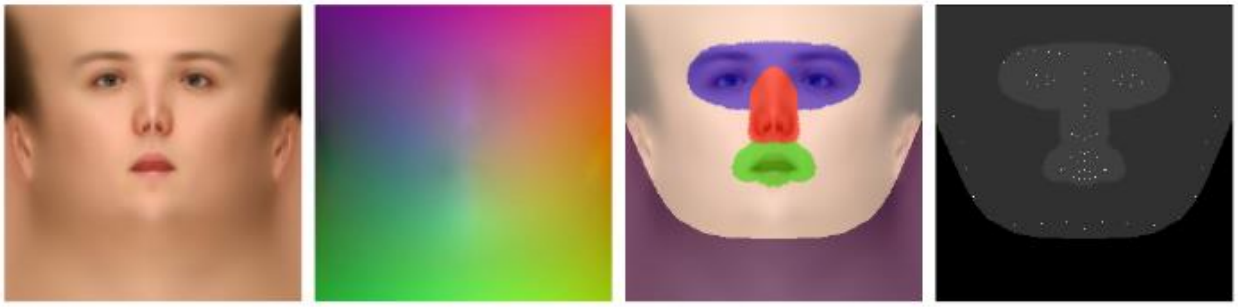
Để học các tham số của mô hình, PRNet xây dựng một hàm mất mát mới để đo sự khác biệt giữa bản đồ vị trí và đầu ra của mô hình. Lỗi bình phương trung bình (MSE) là một hàm mất mát thường được sử dụng cho các bài toán như vậy, chẳng hạn như trong [34, 43]. Tuy nhiên, MSE coi tất cả các điểm như nhau, vì vậy nó không hoàn toàn thích hợp cho việc học bản đồ vị trí. Vì vùng trung tâm của khuôn mặt có nhiều đặc điểm phân biệt hơn các vùng khác, PRNet sử dụng mặt nạ trọng số để tạo lên hàm mất mát.

Như trong hình 2-13, mặt nạ trọng số là một hình ảnh xám ghi lại trọng số của mỗi điểm trên bản đồ vị trí. Nó có cùng kích thước và tỷ lệ với bản đồ vị trí. Theo mục tiêu của PRNet, PRNet phân chia các điểm thành bốn loại, mỗi loại có trọng số riêng trong hàm mất mát. Vị trí của 68 điểm quan trọng trên khuôn mặt có trọng số cao nhất, do đó để đảm bảo mô hình học chính xác vị trí của các điểm này. Vùng cổ thường ít thu hút sự chú ý và thường bị che. Vì việc học hình dạng 3D của cổ hoặc quần áo nằm ngoài mối quan tâm của PRNet, mô hình này chỉ định trọng số 0 cho các điểm ở vùng cổ để giảm bớt nhiễu trong quá trình huấn luyện.

Giả sử biểu thị bản đồ vị trí dự đoán là $P(x, y)$với x, y đại diện cho mỗi tọa độ pixel. Cho bản đồ vị trí ground truth $\tilde{P}(x, y)$và mặt nạ trọng số $W(x, y)$, hàm trọng số được tính bằng:

$$Loss = \sum \|P(x, y) - \tilde{P}(x, y)\| \cdot W(x, y)$$

Cụ thể, PRNet sử dụng tỷ lệ trọng số sau trong các thí nghiệm của mình, tiểu vùng 1 (68 điểm mốc trên khuôn mặt): tiểu vùng 2 (mắt, mũi, miệng): tiểu vùng 3 (vùng mặt khác): tiểu vùng 4 (cổ) = 16: 4: 3: 0. Mặt nạ trọng số cuối cùng được hiển thị trong hình 2-15

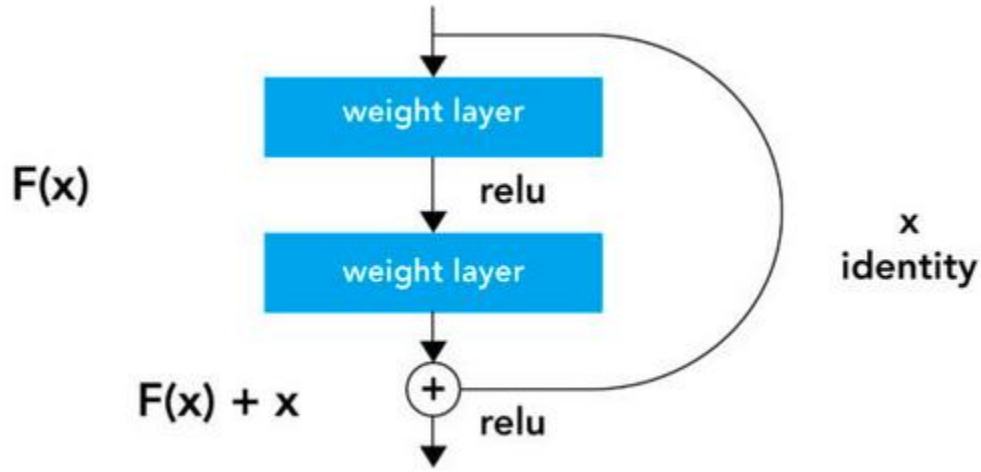


Hình 2-15: Hình minh họa của mặt nạ trọng số. Từ trái sang phải: bản đồ kết cấu UV, bản đồ vị trí UV, bản đồ kết cấu màu với thông tin phân đoạn (màu xanh cho vùng mắt, màu đỏ cho vùng mũi, màu xanh lá cây cho vùng miệng và màu tím cho vùng cổ), mặt nạ trọng số ở cuối cùng.

2.3.3. Mạng ResNet

ResNet, viết tắt của Residual Network là một loại mạng nơ-ron cụ thể đã được giới thiệu vào năm 2015 bởi Kaiming He, Xiangyu Zhang, Shaoqing Ren và Jian Sun.

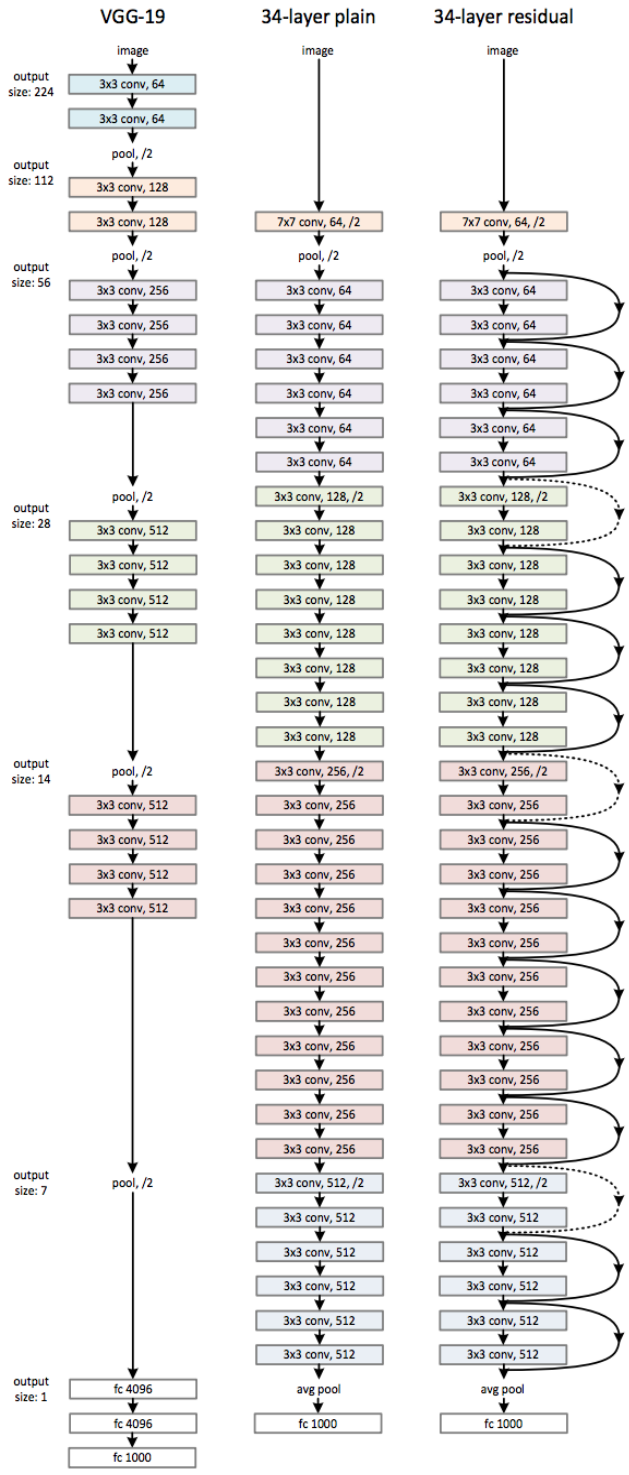
Một vấn đề thường gặp khi huấn luyện một mạng network quá sâu là vanishing problem, và resnet được sinh ra để khắc phục vấn đề này bằng cách đưa ra khối dư thừa (residual block)



Hình 2-16: Khối dư thừa

Sự khác biệt ở khối dư thừa là có một kết nối trực tiếp bỏ qua một số lớp (có thể khác nhau trong các mô hình khác nhau) ở giữa. Kết nối này được gọi là 'kết nối nhảy' (skip connection) và là thành phần cốt lõi của các khối dư thừa. Do kết nối bỏ qua này, đầu ra

của lớp bây giờ không giống nhau. Nếu không sử dụng kết nối nhảy này, đầu vào ‘x’ sẽ được nhân với trọng số của lớp, sau đó thêm một số hạng thiên vị (bias).



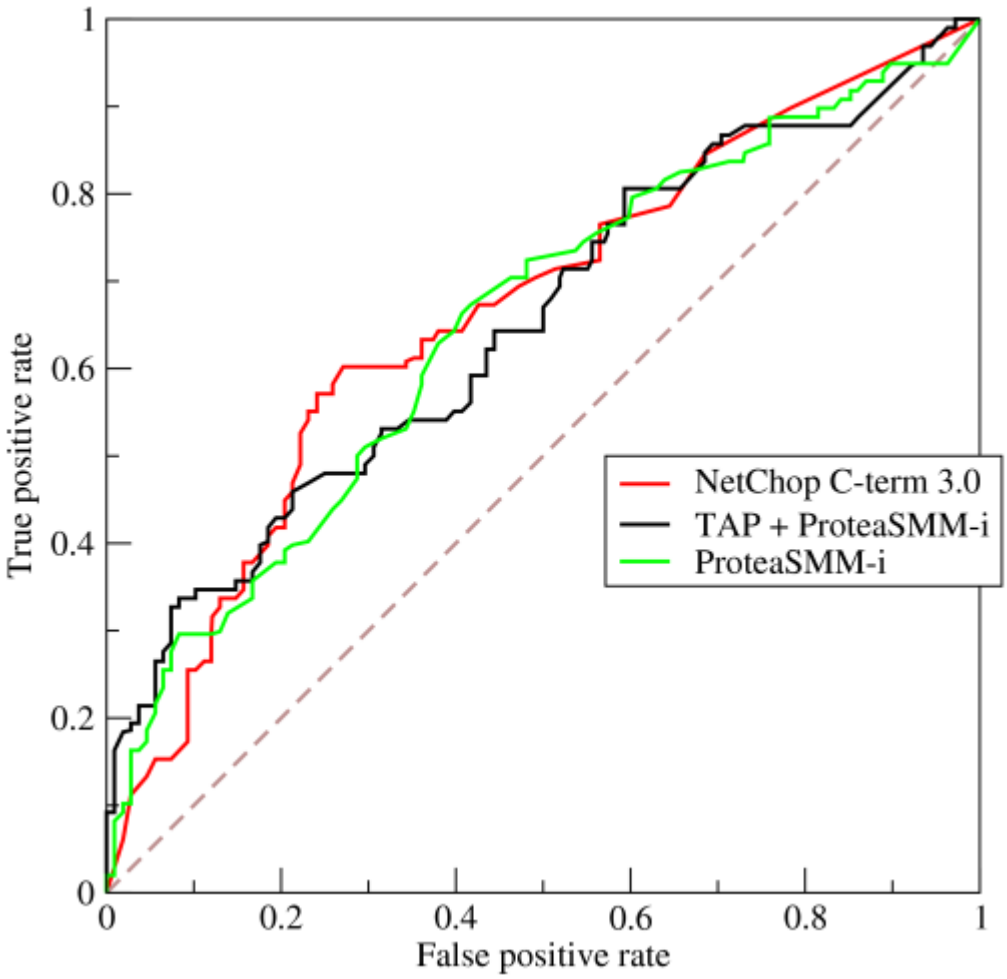
Hình 2-17: So sánh Resnet-34 và một mạng network với 34 lớp

2.3.4. Kết hợp CDC, thông tin chiều sâu và Resnet-34

Như đã mô tả ở các phần trước luận văn sẽ thử nghiệm 3 mô hình cho bài toán phát hiện giả mạo khuôn mặt. Mô hình này kế thừa mô hình trước khi lần lượt ứng dụng thêm vào các kỹ thuật CDC và thông tin về chiều sâu của khuôn mặt. Cụ thể quá trình xây dựng ba mô hình như sau:

- Mô hình 1: Ban đầu mạng resnet-34 có lớp kết nối đầy đủ trả ra 1000 lớp do được huấn luyện với tập dữ liệu ImageNet (Hình 2-17). Tuy nhiên, bài toán phát hiện giả mạo khuôn mặt đang là một bài toán phân lớp có 2 nhãn là khuôn mặt thật và khuôn mặt giả mạo nên luận văn thực hiện thay thế lớp kết nối đầy đủ cuối cùng thành một lớp khác mà đầu ra chỉ trả về 2 giá trị để phù hợp với bài toán. Bên cạnh đó, bởi dữ liệu huấn luyện cho bài toán phát hiện giả mạo đang ít hơn so với tập dữ liệu ImageNet nên luận văn thực hiện đóng băng các lớp trước lớp cuối cùng, hay nói cách khác là không cập nhật trọng số của các lớp này trong quá trình huấn luyện mà chỉ cập nhật trọng số của lớp kết nối đầy đủ cuối cùng.
- Mô hình 2: Ở mô hình số 2, luận văn thực hiện thay thế toàn bộ các lớp tích chập ban đầu có trong mô hình 1 bằng lớp CDC. Ở mô hình này, luận văn tiếp tục thực hiện chiến lược cập nhật trọng số của mô hình số 1, cộng thêm việc cập nhật trọng số của các lớp CDC mới được thêm vào.
- Mô hình 3: Ở mô hình số 3, luận văn tiếp tục thay đổi bắt nguồn từ mô hình số 2. Ở mô hình này luận vặn loại bỏ hẳn lớp kết nối đầy đủ cuối cùng. Sau đó, một lớp Upsample được đặt vào sau khối các lớp thứ 4 (gồm các lớp có số lượng kênh là 512) để thực hiện đưa feature map về kích thước 32x32. Bên cạnh đó, với mỗi một bức ảnh khuôn mặt đầu vào, luận văn sẽ đưa qua PRNet để nhận được một hình ảnh chiều sâu rồi tiếp tục đưa về kích thước 32x32. Sau đó tính tổng các giá trị của 2 ma trận 32x32 rồi tính tỉ lệ, tỉ lệ này được so sánh với một ngưỡng cho trước được tìm từ tập phát triển để kết luận một ảnh là thật hay là giả. Để tìm được ngưỡng này, luận văn thực hiện như sau: Với mỗi ảnh, luận văn luôn tính được tỉ lệ giữa tổng giá trị của bản đồ đặc trưng từ mô hình 3 và tổng giá trị của ma trận

chiều sâu từ PRNet. Như vậy sau khi chạy hết tập phát triển, luận văn có thể thu được một tập các giá trị tỉ lệ này. Bước tiếp theo luận văn đưa tập các giá trị này cùng với nhãn tương ứng của bức ảnh để tạo ra một đường cong Receiver operating characteristic (ROC Curve). Từ ROC curve, một ngưỡng cho việc phân loại khuôn mặt thật và khuôn mặt giả mạo sẽ được lựa chọn.



Hình 2-18: Ví dụ về một đường cong ROC

2.4. Các vấn đề thích ứng miền

Với một tập dữ liệu đã được gắn nhãn với chất lượng tốt, các thuật toán học sâu có thể học để đưa ra các kết quả vô cùng chính xác. Mặt khác, trong trường hợp tập dữ liệu đã được gắn nhãn cho một bài toán cụ thể không đủ lớn thì có một cách khác để xử lý đó là sử dụng một mô hình khác đã được huấn luyện cho một bài toán tương tự. Cách làm này

gọi là học chuyển tiếp. Trong trường hợp này, một số lớp cuối sẽ được tinh chỉnh với tập dữ liệu của bài toán mới cần giải quyết. Tuy nhiên, ở cả hai trường hợp trên thì đều được giả định là dữ liệu huấn luyện ở phân phối cơ bản. Ngược lại, nếu đầu vào ở giai đoạn kiểm thử khác đáng kể so với dữ liệu huấn luyện thì mô hình có thể không còn thực sự rất tốt nữa. Ví dụ với một bài toán phân vùng ảnh, dữ liệu được thu thập từ camera trước của một ô tô để biết các đối tượng nào đang ở phía trước (tòa nhà, cây, người đi bộ, đèn giao thông v.v). Giả sử rằng dữ liệu này được thu thập tại thủ đô Hà Nội và được đưa đi huấn luyện cho một thuật toán phân vùng ảnh, mô hình sau huấn luyện được đưa đi thử nghiệm trên đường phố Hà Nội và cho ra kết quả tốt. Tuy nhiên khi đem mô hình này sang Nhật Bản để thử nghiệm thì kết quả không còn được như cũ nữa do ô tô và cảnh vật ở Nhật Bản khác xa với Hà Nội. Nguyên nhân dẫn tới điều này là miền bài toán đã thay đổi. Cụ thể, miền dữ liệu đầu vào đã thay đổi còn miền bài toán (nhãn) vẫn giữ nguyên. Vấn đề này được gọi là vấn đề về thích ứng miền.

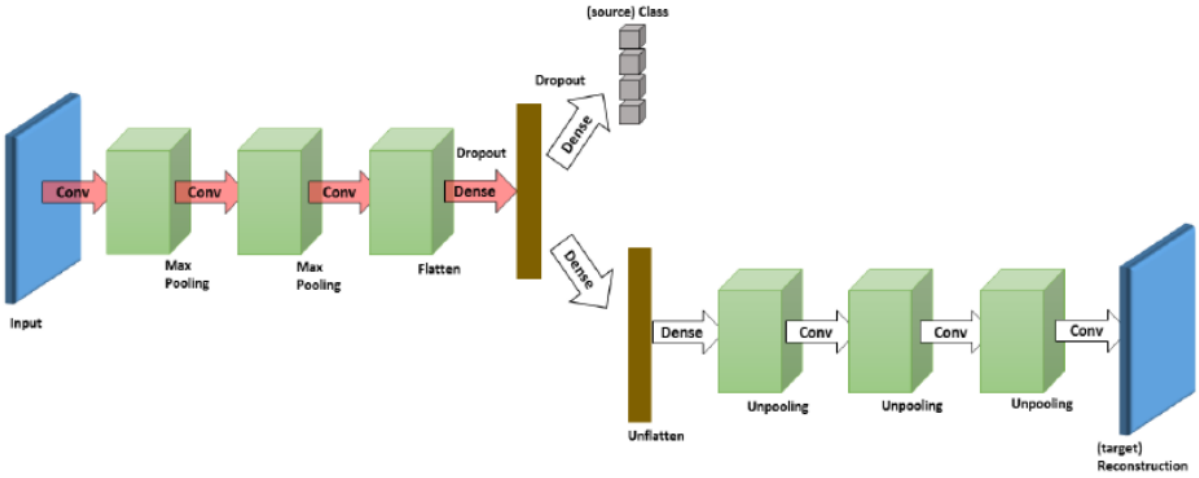
Qua các ví dụ trên, quay lại với bài toán gốc mà luận văn đề cập là phát hiện giả mạo khuôn mặt, ta có thể thấy rằng việc áp dụng các thuật toán học sâu để giải cũng sẽ đối mặt với vấn đề thích ừng miền. Thật vậy, trong thực tế, khi huấn luyện mô hình, ta chưa chắc đã có được nguồn dữ liệu giống với hình thức khi một đối tượng tấn công hệ thống nhận diện khuôn mặt. Bởi thiết bị mà đối tượng sử dụng cho chắc đã giống với thiết bị được dùng để xây dựng bộ dữ liệu, cũng như là hình thức giả mạo mà đối tượng thực hiện cũng chưa chắc đã giống với các cách thức mà bộ dữ liệu huấn luyện đã định nghĩa từ trước. Điều này khiến cho bài toán phát hiện giả mạo khuôn mặt càng trở lên thách thức hơn trong thực tế.

Qua khảo sát từ [47], luận văn thấy rằng có một số hướng tiếp cận để giải quyết vấn đề này đó là:

- Divergence-based domain adapatation: Phương pháp này hoạt động bằng cách giảm thiểu một số tiêu chí phân biệt giữa phân phối của dữ liệu nguồn và phân phối của dữ liệu đích, do đó đạt được các biểu diễn đặc trưng miền bất biến. Nếu có thể tìm thấy được một biểu diễn đặc trưng như vậy, bộ phân loại sẽ có thể hoạt động tốt trên cả hai miền dữ liệu. Tất nhiên, điều này giả định rằng tồn tại một

biểu diễn như vậy, từ đó đưa ra giả định rằng các bài toán có liên quan theo một cách nào đó. Có thể kể đến một số phương pháp theo hướng tiếp cận này như [49, 50, 51, 52].

- Adversarial-based Domain Adaptation: Kỹ thuật này cố gắng đạt được sự thích ứng miền bằng cách sử dụng huấn luyện đối nghịch. Một cách tiếp cận là tạo dữ liệu mục tiêu (target dataset) có liên quan đến miền nguồn (source domain) (ví dụ giữ nguyên các nhãn) bằng cách sử dụng Generative adversarial networks (GAN) [53]. Những dữ liệu mới được tổng hợp sau đó sẽ được sử dụng để huấn luyện mô hình cho tập dữ liệu mục tiêu.
- Reconstruction-based Domain Adaptation: Cách tiếp cận này sử dụng một nhiệm vụ tái thiết bổ trợ (auxiliary reconstruction task) để tạo ra một biểu diễn có thể dùng chung cho mỗi miền. Ví dụ, phương pháp [54] đề cập tới một luồng xử lý để học cách chuyển đổi hình ảnh nguồn thành hình ảnh giống với tập dữ liệu đích.

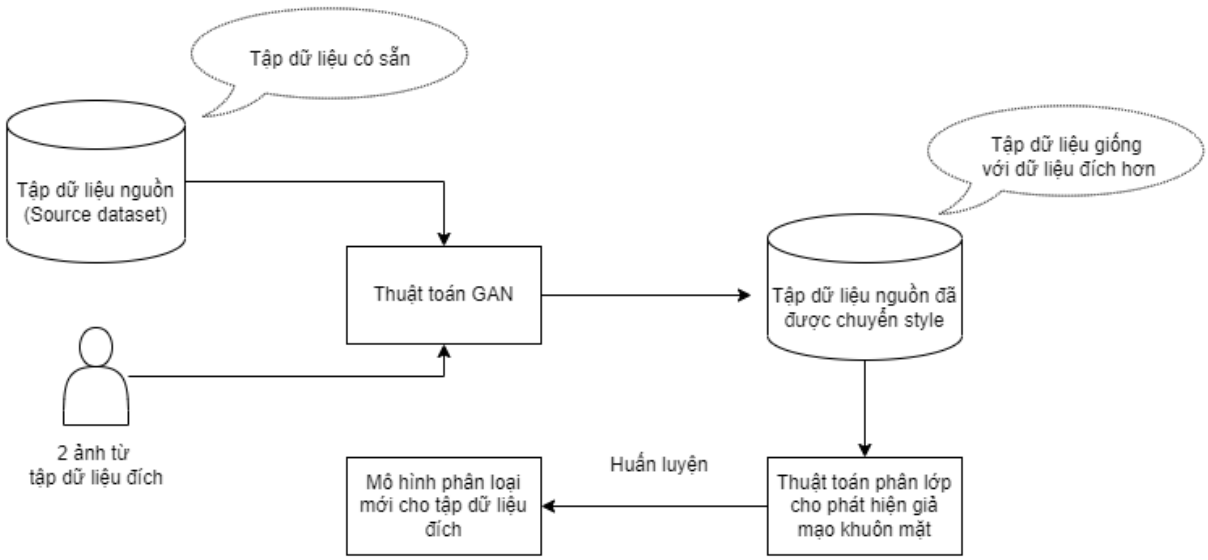


Hình 2-19: Kiến trúc của phương pháp [54]

2.5. Ứng dụng GAN cho vấn đề thích ứng miền

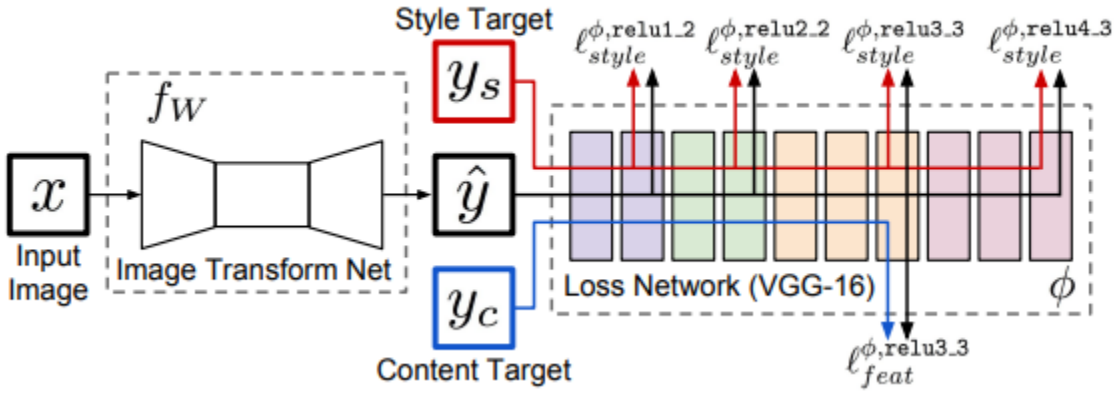
Trong quá trình thực hiện, luận văn tiếp cận theo hướng Adversarial-based Domain Adaptation để khắc phục vấn đề thích ứng miền. Cụ thể, luận văn thực hiện lấy một số hình ảnh ở tập dữ liệu đích làm hình ảnh mẫu để một mô hình chuyển kiểu (transfer style) học, sau khi kết thúc quá trình học, mô hình chuyển kiểu này sẽ được sử dụng để chuyển đổi

tập dữ liệu gốc mà ta đang sẵn có, sau đó tập dữ liệu mới này được dùng để huấn luyện lại một mô hình phát hiện giả mạo khuôn mặt mới (Hình 2-20).



Hình 2-20: Luồng thực hiện GAN trong luận văn

Để thực hiện ý tưởng này, luận văn có sử dụng một thuật toán chuyển kiểu được mô tả trong bài báo "Perceptual Losses for Real-Time Style Transfer and Super-Resolution" [55].



Hình 2-21: Kiến trúc tổng quan mạng chuyển đổi kiểu [55]

Kiến trúc được sử dụng trong bài báo [55] gồm 2 thành phần (Hình 2-21): một mạng chuyển đổi hình ảnh (image transformation network) f_w và một mạng mất mát (loss network) Ø được sử dụng để định nghĩa các hàm mất mát $l_1, l_2, \ldots, l_k$. Mạng chuyển đổi hình ảnh ở đây thực chất là một mạng tích chập dư thừa sâu (deep residual convolutional

neural network) tham số hóa bởi trọng số W. Với một bức ảnh x sẽ được chuyển đổi thành $\hat{y} = f_w(x)$. Mỗi hàm mất mát sẽ tính ra một giá trị mất mát $l_i(\hat{y}, y_i)$ để đo lường sự khác biệt giữa ảnh đầu ra $\hat{y}$ và ảnh mục tiêu y_i. Mạng chuyển đổi hình ảnh được huấn luyện bằng thuật toán stochastic gradient descent để cực tiểu hóa một hàm tổng hợp đánh trọng số cho các hàm mất mát:

$$W^* = argmin_W E_{x,\{y_i\}}[\sum_{i=1} \lambda_i l_i(f_W(x), y_i)]$$

Ở thành phần thứ hai là mạng mất mát $\emptyset$, $\emptyset$ sẽ là một mạng đã được tiền huấn luyện cho bài toán phân loại hình ảnh hay nói cách khác mạng mất mát được cố định trong thuật toán để định nghĩa ra các hàm mất mát. Các hàm mất mát được định nghĩa tư mạng mất mát này gồm hàm mất mát tái cấu tạo đặc trưng (feature reconstruction loss) $l^{\emptyset}_{feat}$ và hàm mất mát tái cấu tạo kiểu (style reconstruction loss) $l^{\emptyset}_{style}$ để đo lường sự khác biệt ở nội dung và kiểu giữa các hình ảnh. Cụ thể, mục tiêu nội dung y_c chính là ảnh đầu vào x, và mục tiêu kiểu là y_s. Hình ảnh đầu ra cần phải kết hợp được hai hình ảnh mục tiêu này.

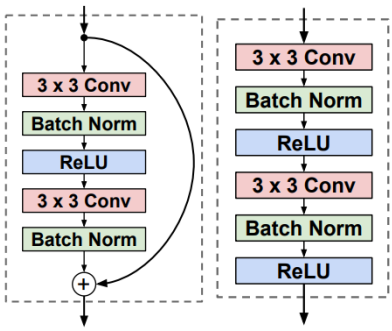
2.5.1. Mạng chuyển đổi hình ảnh

Kiến trúc của mạng chuyển đổi hình ảnh không sử dụng bất kỳ lớp pooling nào mà sử dụng các lớp tích chập với thông số về bước nhảy (stride) khác nhau để thực hiện giảm kích thước (downsampling) và tăng kích thước (upsampling). Cụ thể, mạng gồm 5 khối dư thừa (residual block) [42] sử dụng theo kiến trúc của [56]. Tất cả các lớp tích chập không dư thừa (non-residual) được theo sau bởi spatial batch normalization [57] và ReLU không tuyến tính ngoại trừ lớp đầu ra, mà thay vào đó là scaled tanh để đảm bảo rằng ảnh đầu ra có giá trị pixel nằm trong khoảng từ 0 đến 255. Khác với lớp đầu tiên và lớp cuối cùng sử dụng kernel kích thước 9x9, tất cả các lớp tích chập sử dụng kích thước kernel là 3x3 (Hình 2-22)

Layer	Activation size
Input	$3 \times 256 \times 256$
$32 \times 9 \times 9$ conv, stride 1	$32 \times 256 \times 256$
$64 \times 3 \times 3$ conv, stride 2	$64 \times 128 \times 128$
$128 \times 3 \times 3$ conv, stride 2	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
Residual block, 128 filters	$128 \times 64 \times 64$
$64 \times 3 \times 3$ conv, stride 1/2	$64 \times 128 \times 128$
$32 \times 3 \times 3$ conv, stride 1/2	$32 \times 256 \times 256$
$3 \times 9 \times 9$ conv, stride 1	$3 \times 256 \times 256$

Hình 2-22: Kiến trúc của mạng chuyển đổi kiểu

Bài báo [42] có đưa ra quan điểm rằng các kết nối dư thừa có tác dụng giúp cho mạng có thể học được các hàm nhận dạng (identity function) dễ dàng hơn, trong khi đó trong hầu hết các trường hợp của bài toán chuyển kiểu thì ảnh đầu ra cần có điểm chung về kiến trúc với ảnh đầu vào, vì vậy các kết nói dư thừa có thể phù hợp với mục tiêu của mạng chuyển đổi hình ảnh. Bởi lẽ đó mà trong kiến trúc của mạng chuyển đổi hình ảnh có chứa một vài khối dư thừa (residual block) (Hình 2-23)



Hình 2-23: Bên trái là khối dư thừa được sử dụng, bên phải là khối tích chập thông thường

2.5.2. Hàm mất mát tri giác (Perceptual Loss function)

Phương pháp [55] định nghĩa hai hàm mất mát tri giác đo lường sự khác biệt về tri giác và ngữ nghĩa (semantic) của các hình ảnh ở mức cao. Trong đó mạng mất mát $Ø$ được tiền huấn luyện cho phân loại hình ảnh, cụ thể ở đây là mạng VGG-16 [59] được tiền huấn luyện trên bộ dữ liệu ImageNet.

Hàm mất mát tri giác đầu tiên đó là hàm mất mát tái cấu tạo đặc trưng (feature reconstruction loss). Thay vì mục tiêu là các pixel của ảnh đầu ra $\hat{y}$ giống hệt với các pixel của ảnh mục tiêu y thì ở phương pháp này sẽ thúc đẩy để các bức ảnh này tương tự nhau ở biểu diễn đặc trưng khi được tính bởi mạng mất mát $Ø$. Gọi $Ø_j(x)$ là kích hoạt (activations) của lớp thứ j của mạng $Ø$ khi xử lý ảnh x. Nếu j là một lớp tích chập thì $Ø_j(x)$ sẽ là một bản đồ đặc trưng có kích thước $C_j \; x \; H_j \; x \; W_j$. Hàm mất mát tái cấu tạo đặc trưng là khoảng cách Euclidean giữa các biểu diễn đặc trưng được tính theo công thức:

$$l_{feat}^{Ø,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \left\| Ø_j(\hat{y}) - Ø_j(y) \right\|_2^2$$

Việc sử dụng một hàm mất mát tái cấu tạo đặc trưng cho quá trình huấn luyện mạng chuyển kiểu ảnh sẽ giúp ảnh đầu ra $\hat{y}$ tương tự với ảnh đầu vào y nhưng không bắt buộc 2 bức ảnh là giống nhau chính xác hoàn toàn.

Hàm mất mát tri giác thứ hai đó là hàm mất mát tái cấu tạo kiểu (style reconstruction loss). Mục đích của hàm mất mát này sẽ là đo lường sự khác biệt giữa hai bức ảnh ở các đặc điểm như màu, cấu trúc, mẫu chung (common patterns), v.v.

Đầu tiên, vẫn giữ nguyên định nghĩa của $Ø_j(x)$ như phía trên. Gọi ma trận Gram $G_j^{Ø}(x)$ là ma trận có kích thước $C_j \; x \; C_j$, tính bởi công thức sau:

$$G_j^{Ø}(x)_{c,c'} = \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} Ø_j(x)_{h,w,c} \, Ø_j(x)_{h,w,c'}$$

Hàm mất mát tái cấu tạo kiểu sẽ là squared Frobenius norm của sự khác biệt giữa ma trận Gram của ảnh đầu ra và ảnh mục tiêu:

$$l_{style}^{Ø,j}(\hat{y}, y) = \left\| G_j^{Ø}(\hat{y}) - G_j^{Ø}(y) \right\|_F^2$$

Việc tạo một bức ảnh $\hat{y}$ cực tiểu hóa hàm mất mát tái cấu tạo kiểu sẽ giữ lại các đặc trưng về kiểu cách từ ảnh mục tiêu nhưng không giữ lại cấu trúc không gian của nó. Bên cạnh đó để thực hiện tái cấu tạo kiểu từ một tập các lớp J thay vì một lớp j đơn lẻ, bài báo định nghĩa $l_{style}^{\emptyset,J}(\hat{y}, y)$ là tổng của các hàm mất mát của các lớp $j \in J$.

2.6. Kết luận

Chương này đã trình bày các cơ sở lý thuyết của Central Difference Convolution, dựng thông tin chiều sâu cho khuôn mặt, residual network, và vấn đề thích ứng miền có thể gặp phải trong bài toán phát hiện giả mạo khuôn mặt và ý tưởng nhằm khắc phục vấn đề này dựa trên [55].

Chương tiếp theo sẽ mô tả quá trình chuyển bị dữ liệu, các độ đo nhằm đánh giá giải pháp cùng các thử nghiệm được thực hiện và kết quả cuối cùng.

CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ

Chương 3 sẽ trình bày về các tập dữ liệu được sử dụng trong luận văn cũng như các bước tiền xử lý dữ liệu trong quá trình thực hiện, mô tả rõ hơn các thông số trong từng thực nghiệm như đã giới thiệu ở chương 2 cùng kết quả của các thực nghiệm đó, cuối cùng từ những kết quả này đưa ra một số nhận xét.

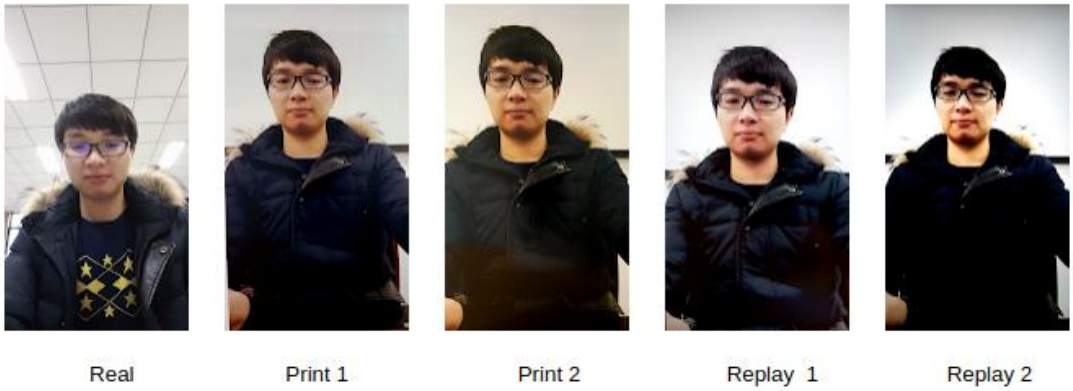
3.1. Dữ liệu thử nghiệm

3.1.1. Tập dữ liệu OULU

Tập dữ liệu phát hiện giả mạo khuôn mặt của Oulu-NPU bao gồm 4950 video về các cuộc tấn công, giả mạo và khuôn mặt thật. Các video này được quay bằng camera trước của sáu thiết bị di động (Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, ASUS Zenfone Selfie, Sony XPERIA C5 Ultra Dual và OPPO N3) trong ba khoảng thời gian với các điều kiện ánh sáng và cảnh nền khác nhau (Khoảng thời gian 1, Khoảng thời gian 2 và khoảng thời gian 3). Các kiểu tấn công được xem xét trong tập dữ liệu OULU-NPU là in (printer) và phát lại video (replay attack). Các cuộc tấn công được tạo ra bằng cách sử dụng hai máy in (Máy in 1 và Máy in 2) và hai thiết bị hiển thị (Màn hình 1 và Màn hình 2). Hình 3-1 cho thấy một số hình ảnh mẫu về các cuộc tấn công và truy cập thực được chụp bằng điện thoại Samsung Galaxy S6 edge. Các video của 55 đối tượng được chia thành ba tập con tách rời để huấn luyện, phát triển và kiểm tra. Bảng sau đây cung cấp tổng quan chi tiết về sự phân chia trong tập dữ liệu này.

Bảng 3-1: Thông tin chi tiết về tập dữ liệu

	Users	Real access	Print attacks	Video attacks	Total
Training	20	360	720	720	1800
Development	15	270	540	540	1350
Test	20	360	720	720	1800



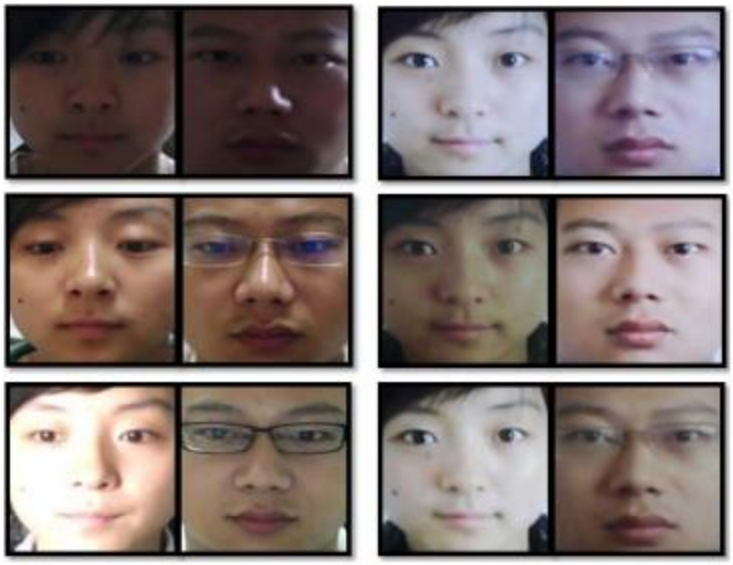
Hình 3-1: Một số hình ảnh ví dụ về khuôn mặt thật và các khuôn mặt giả mạo

Trong quá trình thực hiện, do tập dữ liệu OULU quá lớn (hơn 100Gb) nên luận văn đã thực hiện thí nghiệm trên một tập con của tập dữ liệu gốc. Cụ thể với mỗi video gốc, luận văn thực hiện tách từng khung hình và sau đó giữ lại ngẫu nhiên một khung hình duy nhất. Tổng dung lượng sau khi rút gọn là 1Gb. Sau khi chia tập dữ liệu theo một cách chia sẵn có của OULU, luận văn sử dụng 1200 ảnh cho tập huấn luyện (training), 900 ảnh cho tập phát triển (development), 600 ảnh cho tập kiểm thử (testing).

3.1.2. Tập dữ liệu NUAA

Tập dữ liệu NUAA được xây dựng bằng cách sử dụng một webcam giá rẻ thông thường và được thu thập trong 3 khoảng thời gian khác nhau, mỗi khoảng thời gian cách 2 tuần, địa điểm và điều kiện chiếu sáng của khoảng thời gian cũng khác nhau. Có tổng cộng 15 đối tượng (đánh số từ 1 đến 15) xuất hiện trong tập dữ liệu này. Trong mỗi khoảng thời gian, tập dữ liệu chứa cả ảnh chụp trực tiếp và ảnh chụp gián tiếp của các đối tượng. Một số hình ảnh mẫu từ khoảng thời gian được đưa ra trong hình 3-2. Cụ thể, đối với mỗi đối tượng trong mỗi khoảng thời gian, webcam được sử dụng để chụp một loạt ảnh khuôn mặt của họ (với tốc độ khung hình 20fps và 500 ảnh cho mỗi đối tượng).

Khi chụp ảnh, mỗi đối tượng được yêu cầu nhìn vào webcam trực diện và với biểu cảm trung tính và không có chuyển động rõ ràng như chuyển động của mắt hoặc đầu. Nói cách khác, một đối tượng được yêu cầu làm cho càng giống một bức ảnh càng tốt (ngược lại đối với ảnh chụp). Một số ví dụ về các hình ảnh đã chụp được minh họa trong hình 3-2 (cột bên trái).



Hình 3-2: Hình minh họa từ tập dữ liệu. Trong mỗi cột (từ trên xuống dưới) các mẫu lần lượt là từ khoảng thời gian 1, khoảng thời gian 2 và khoảng thời gian 3. Trong mỗi hàng, cặp bên trái là hình ảnh từ khuôn mặt thật và cặp bên phải là khuôn mặt giả mạo. Lưu ý rằng nó chứa các thay đổi ngoại hình khác nhau mà hệ thống nhận dạng khuôn mặt thường gặp phải (ví dụ: giới tính, khả năng chiếu sáng, có / không đeo kính). Tất cả ảnh gốc trong cơ sở dữ liệu đều là ảnh màu có cùng độ phân giải 640 × 480 pixel.

Trong quá trình xây dựng bộ dữ liệu, mỗi đối tượng được chụp ảnh độ nét cao bằng máy ảnh Canon thông thường đảm bảo vùng mặt phải chiếm ít nhất 2/3 diện tích toàn bức ảnh. Sau đó, hình ảnh giả mạo được tạo bằng 2 cách. Đầu tiên là sử dụng phương pháp truyền thống để in những bức ảnh này trên giấy ảnh với kích thước phổ biến lần lượt là 6,8cm × 10,2cm (nhỏ) và 8,9cm × 12,7cm (lớn hơn). Cách thứ 2, mỗi bức ảnh chất lượng cao được in trên giấy A4 70g bằng máy in HP màu thông thường. Dựa trên những điều này, ba loại tấn công ảnh được mô phỏng trước webcam như trong hình 3-3.

Hình 3-3: Hình minh họa các cuộc tấn công ảnh khác nhau (từ trái sang phải): (1) di chuyển ảnh theo chiều ngang, theo chiều dọc, phía sau và phía trước; (2) xoay ảnh theo chiều sâu dọc theo trục dọc; (3) giống với (2) nhưng dọc theo trục hoành; (4) bẻ cong ảnh vào trong và ra ngoài theo trục tung; (5) giống như (4) nhưng dọc theo trục hoành.

3.2. Các độ đo

Để có thể đánh giá được các thuật toán, mô hình học sâu được thử nghiệm thì không thể thiếu đi được các độ đo phù hợp với bài toán. Cụ thể trong bài toán phát hiện giả mạo khuôn mặt, luận văn sử dụng 3 độ đo chính là APCER, BPCER và ACER.

Để có thể hiểu rõ hơn về các độ đo này, luận văn sẽ định nghĩa về các khái niệm True Positive (TP), True Negative (TN), False Positive (FP) và False Negative (FN).

Cụ thể:

TP: Là số khuôn mặt giả mạo được nhận đúng khuôn mặt giả mạo;

TN: Là số khuôn mặt thật được nhận đúng là khuôn mặt thật;

FP: Là số khuôn mặt thật bị nhận sai là khuôn mặt giả mạo;

FN: Là số khuôn mặt giả mạo bị nhận sai là khuôn mặt thật;

Các độ đo được dùng trong luận văn được định nghĩa cụ thể như sau:

- Attack Presentation Classification Error Rate (APCER)

  $$APCER = FN / (TP + FN)$$

- Bona Fide Presentation Classification Error Rate (BPCER):

  BPCER= FP / (FP + TN)

- Average Classification Error Rate (ACER):

  ACER = (APCER + NPCER) / 2

3.3. Thực nghiệm

3.3.1. Thực nghiệm với riêng mạng resnet-34

Ở thực nghiệm này, mô hình đang dùng để huấn luyện là resnet-34 giữ nguyên từ bài báo gốc, trừ lớp kết nối đầy đủ được thay đổi để trả về phân bố xác suất của 2 nhãn là khuôn mặt giả mạo và khuôn mặt thật.

Ở thí nghiệm này các thông số cho quá trình huấn luyện như sau:

- Số lượng epoch: 300
- Optimizer là Adam với learning rate là 0.001
- Loss function là Cross Entropy
- Pre-train model được huấn luyện từ tập dữ liệu ImageNet

3.3.2. Thực nghiệm với mạng resnet-34 kết hợp CDC

Ở thực nghiệm này, các lớp tích chập trong mạng resnet-34 đã được thay thế bởi CDC, các thông số về số lượng epoch, optimizer, và loss function giống với ở mục 3.3.1. Thông số theta cho CDC được chọn là 0.7. Các thông số về in channel, out channel, kernel size, stride, padding được giữ nguyên của các lớp tích chập tương ứng đã được thay thế. Dưới đây là phần mã nguồn của CDC sử dụng PyTorch.

```python
class Conv2d_cd(nn.Module):
    def __init__(self, in_channels, out_channels, kernel_size=3, stride=1,
                 padding=1, dilation=1, groups=1, bias=False, theta=0.7):

        super(Conv2d_cd, self).__init__()
        self.conv = nn.Conv2d(in_channels, out_channels, kernel_size=kernel_size, stride=strid
                              dilation=dilation, groups=groups, bias=bias)
        self.theta = theta

    def forward(self, x):
        out_normal = self.conv(x)

        if math.fabs(self.theta - 0.0) < 1e-8:
            return out_normal
        else:
            # pdb.set_trace()
            [C_out, C_in, kernel_size, kernel_size] = self.conv.weight.shape
            kernel_diff = self.conv.weight.sum(2).sum(2)
            kernel_diff = kernel_diff[:, :, None, None]
            out_diff = F.conv2d(input=x, weight=kernel_diff, bias=self.conv.bias, stride=self.
                                groups=self.conv.groups)

            return out_normal - self.theta * out_diff
```

Để thay thế toàn bộ lớp tích chập bằng CDC, luận văn thực hiện duyệt toàn bộ mô hình, khi phát hiện một lớp là conv (tích chập) sẽ thực hiện thay thế bằng CDC. Đoạn mã nguồn dưới đây thực hiện quá trình này.

```python
def replace_layers(model, old):
    for n, module in model.named_children():
        if len(list(module.children())) > 0:
            replace_layers(module, old)

        if isinstance(module, old):
            new = Conv2d_cd(in_channels=module.in_channels,
                            out_channels=module.out_channels,
                            kernel_size=module.kernel_size[0],
                            stride=module.stride[0],
                            padding=module.padding[0],
                            bias=False,
                            theta=0.7)

            setattr(model, n, new)
```

Ở các thử nghiệm trên, luận văn sử dụng các phép biến đổi ảnh được liệt kê trong bảng dưới đây, các phép biến đổi được lựa chọn có tham khảo từ các mô hình phân loại ảnh được giới thiệu bởi PyTorch.

Bảng 3-1: Các phép biến đổi ảnh ở thí nghiệm resnet-34 cùng CDC

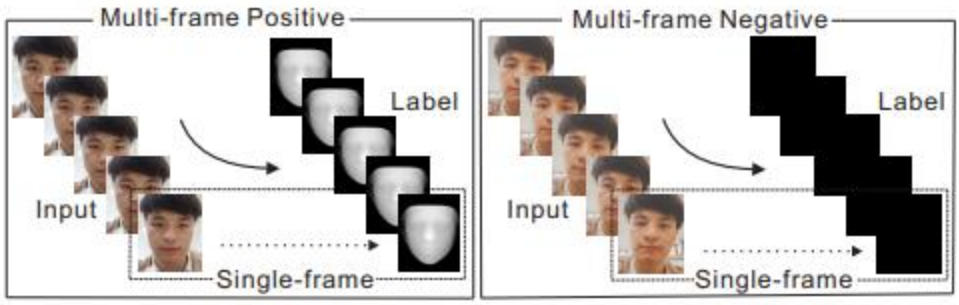
Phép biến đổi	Chú giải
RandomResizedCrop	Chọn ngẫu nhiên một vùng để crop và resize về kích thước phù hợp với đầu vào của resnet-34
RandomHorizontalFlip	Lật ngang bức ảnh ngẫu nhiên
Normalization	Chuẩn hóa giá trị của ảnh về khoảng [0, 1] với mean = [0.485, 0.456, 0.406] và std = [0.229, 0.224, 0.225]

3.3.3. Thực nghiệm với mạng resnet-34 kết hợp CDC và thông tin chiều sâu

Ở thực nghiệm này, các lớp tích chập tiếp tục được thay thế bởi CDC, tuy nhiên lớp kết nối đầy đủ cuối cùng được loại bỏ, thay vào đó là một lớp upsampling để đưa kích thước của feature map cuối cùng trở về 32x32.

Do quá trình sinh ra một ảnh chiều sâu cần nhiều thời gian, do vậy luận văn thực hiện sinh tất cả các hình ảnh chiều sâu cho từng khuôn mặt trước. Cụ thể, với mỗi bức ảnh luận văn sử dụng Multi-task Cascaded Convolutional Networks (MTCNN) để phát hiện khuôn mặt trong từng bức ảnh đó, tiếp theo sử dụng PRNet để sinh ra thông tin chiều sâu của khuôn mặt vừa được phát hiện.

Trong quá trình huấn luyện, các hình ảnh chiều sâu này cũng sẽ được đọc cùng lúc với hình ảnh gốc. Tuy nhiên, đối với các hình ảnh giả mạo, luận văn sẽ thực hiện khởi tạo lại ma trận chiều sâu với toàn bộ các giá trị trong ma trận là 0 (Hình 3-4)



Hình 3-4: Ví dụ về việc tạo ma trận chiều sâu trong quá trình huấn luyện

Bên cạnh đó, luận văn cũng sử dụng thêm một số phép biến đổi để làm đa dạng thêm dữ liệu huấn luyện. Các phép biến đổi này được chọn tương tự như trong bài báo [44] được liệt kê tại bảng sau:

Bảng 3-2: Các phép biến đổi ảnh của thí nghiệm resnet-34, CDC và thông tin chiều sâu

Phép biến đổi	Chú giải
RandomErasing	Random Erasing Data Augmentation bởi Zhong và cộng sự [46]
Cutout	Chọn ngẫu nhiên một khu vực trên ảnh và xóa đen toàn bộ khu vực đó
RandomHorizontalFlip	Lật ngang bức ảnh ngẫu nhiên
Normalization	Chuẩn hóa giá trị của ảnh về khoảng [-1, 1], đối với ảnh chiều sâu là khoảng giá trị [0, 1]

Thông số cụ thể cho quá trình huấn luyện như sau:

- Số lượng epoch: 300
- Optimizer là Adam với learning rate là 0.0001, weight_decay là 5e-4
- Loss function là absolute loss và contrastive loss

Trong đó, absolute loss chính là hàm mất mát MSE. Hàm mất mát này đo lường sai số bình phương trung bình (chuẩn L2 bình phương) giữa mỗi phần tử trong đầu vào x và mục tiêu y.

Với đặc thù là phát hiện giả mạo, hàm mất mát contrastive được cho là giúp kiến trúc học sâu học được các chi tiết tốt hơn [45].

Cụ thể hàm mất mát contrastive được tính dựa trên công thức sau:

$$\mathrm{L}_{single}^{contrast} = \sum_{i} ||\mathrm{K}_i^{contrast} \odot \mathrm{D}_{single} - \mathrm{K}_i^{contrast} \odot \mathrm{D}||_2^2,$$

Với D_{single} là ma trận chiều sâu sinh ra bởi mạng học sâu sử dụng trong luận văn, D là ma trận chiều sâu sinh ra bởi PRNet. $K_i^{contrast}$ là các kernel tích chập tương phản (contrastive convolution kernel) được liệt kê cụ thể dưới đây.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_0, \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_1, \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_2, \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_3,$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}_4, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}_5, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}_6, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_7$$

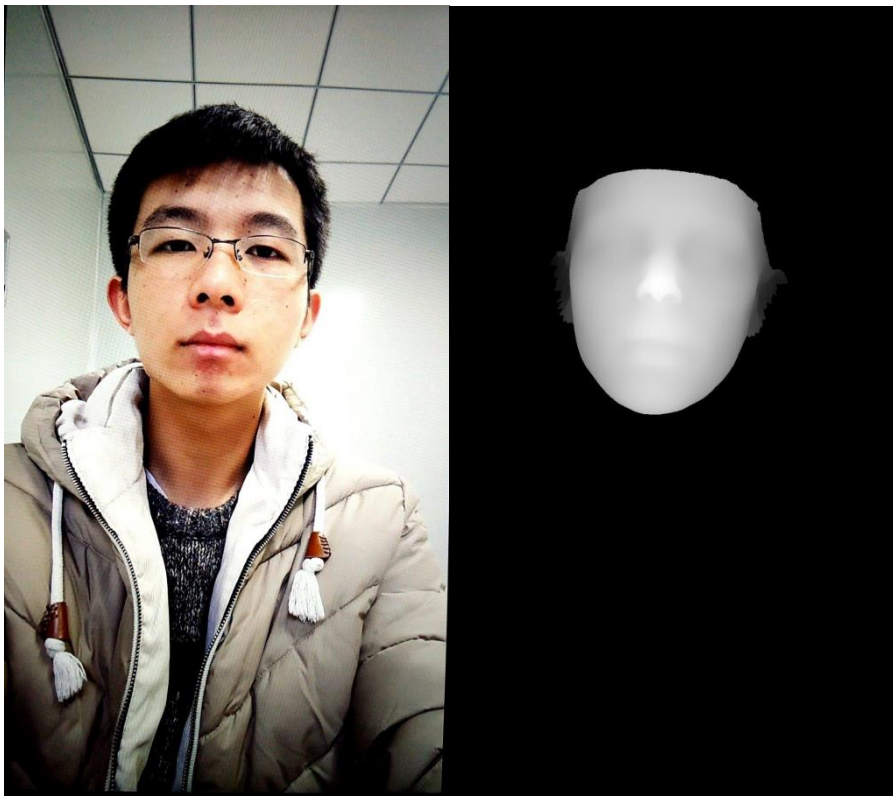
Như công thức của hàm mất mát contrastive, ta có thể thấy được rằng hàm mất mát này nhằm mục đích học đặc điểm về vị trí, địa hình của mỗi pixel, từ đó đưa ra các ràng buộc về độ tương phản từ pixel với các điểm ảnh lân cận.

Cuối cùng, hàm mất mát sử dụng cuối là tổng của absolute loss và contrastive loss.

Dưới đây là phần mã nguồn cho hàm mất mát contrastive.

```python
def contrast_depth_conv(input, device):
    """
    compute contrast depth in both of (out, label)
    :param input: 32x32
    :return: 8 x 32 x 32
    """

    kernel_filter_list = [
        [[1, 0, 0], [0, -1, 0], [0, 0, 0]], [[0, 1, 0], [0, -1, 0], [0, 0, 0]],
[[0, 0, 1], [0, -1, 0], [0, 0, 0]],
        [[0, 0, 0], [1, -1, 0], [0, 0, 0]], [[0, 0, 0], [0, -1, 1], [0, 0, 0]],
        [[0, 0, 0], [0, -1, 0], [1, 0, 0]], [[0, 0, 0], [0, -1, 0], [0, 1, 0]],
[[0, 0, 0], [0, -1, 0], [0, 0, 1]]
    ]

    kernel_filter = np.array(kernel_filter_list, np.float32)

    kernel_filter = torch.from_numpy(kernel_filter.astype(np.float)).float()
    if device.type == 'cuda':
        kernel_filter = kernel_filter.cuda()
    # weights (in_channel, out_channel, kernel, kernel)
    kernel_filter = kernel_filter.unsqueeze(dim=1)

    input = input.unsqueeze(dim=1).expand(input.shape[0], 8, input.shape[1], input.shape[2])

    contrast_depth = F.conv2d(input, weight=kernel_filter, groups=8)  # depthwise conv

    return contrast_depth


class ContrastDepthLoss(nn.Module):
    def __init__(self, device):
        super(ContrastDepthLoss, self).__init__()
        self.device = device
        return

    def forward(self, out, label):
        """
        compute contrast depth in both of (out, label),
        then get the loss of them
        tf.atrous_convd match tf-versions: 1.4
        """
        contrast_out = contrast_depth_conv(out, self.device)
        contrast_label = contrast_depth_conv(label, self.device)

        criterion_MSE = nn.MSELoss().to(device=self.device)

        loss = criterion_MSE(contrast_out, contrast_label)
        # loss = torch.pow(contrast_out - contrast_label, 2)
        # loss = torch.mean(loss)

        return loss
```

Đây là hình ảnh ví dụ chứa thông tin chiều sâu được dựng từ PRNet ở quá trình tiền chuẩn bị dữ liệu:



Hình 3-5: Ví dụ về ảnh chiều sâu được dựng từ PRNet

3.3.4. So sánh các kết quả thử nghiệm

Sau khi thực nghiệm với 3 phương án như trên, luận văn thu được kết quả sau:

Bảng 3-3: Kết quả thử nghiệm

	APCER	BPCER	ACER
Resnet-34	0.0958	0.175	0.1354
CDC + resnet-34	0.0313	0.266	0.1489
CDC + resnet-34 + thông tin chiều sâu	0.019	0.85	0.43

Qua kết quả trên, với độ đo APCER tỉ lệ lỗi giảm dần khi lần lượt kết hợp resnet-34 với CDC và ảnh chiều sâu, điều này cho thấy rằng CDC cùng với thông tin về chiều sâu có

tác động tốt trong việc phát hiện được các trường hợp giả mạo, khi tỷ lệ phát hiện nhầm khuôn mặt giả mạo thành khuôn mặt thật là thấp.

Tuy nhiên điều ngược lại xuất hiện ở độ đo BPCER khi tỷ lệ lỗi lại tăng dần, đặc biệt CDC + resnet-34 + thông tin chiều sâu lại có tỉ lệ lỗi lớn hơn khá nhiều so với 2 thử nghiệm còn lại.

Với độ đo ACER là trung bình cộng của hai độ đo trên thì resnet-34 có nhỉnh hơn một chút so với CDC + resnet-34 tuy nhiên với mục đích là phát hiện khuôn mặt giả mạo mà vẫn cân bằng được 2 độ đo APCER và BPCER thì sự kết hợp giữa CDC và resnet-34 đang cho thấy những ưu thế của phương pháp này.

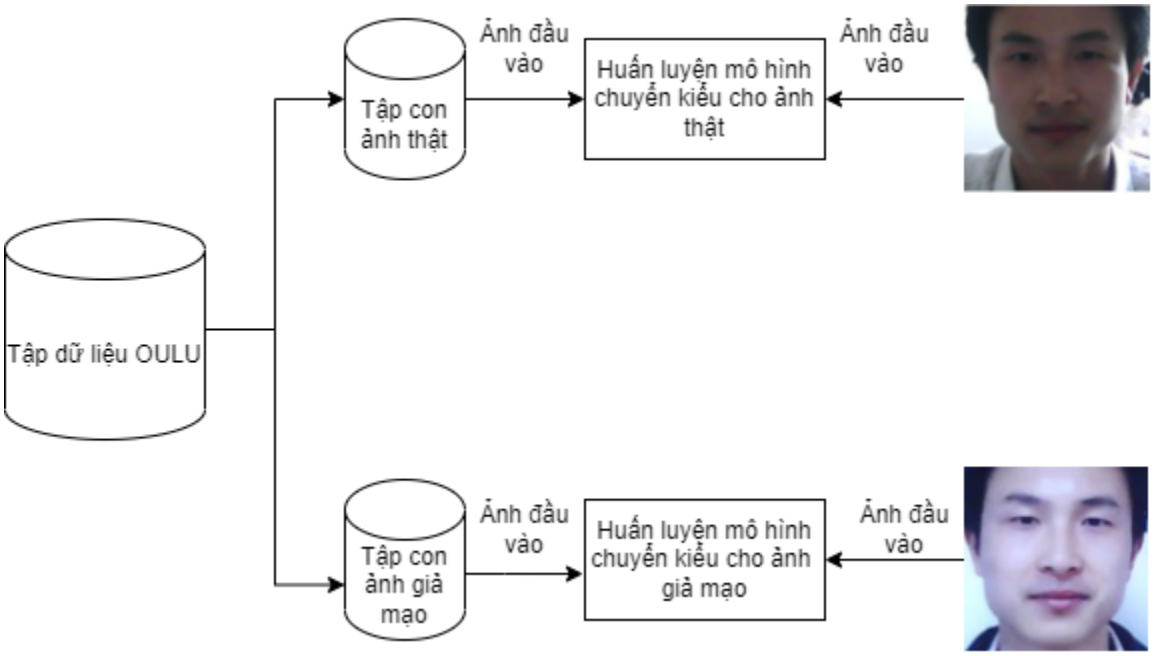
3.4. Thử nghiệm GAN trong vấn đề thích ứng miền

Như đã trình bày tại mục 2.5, để huấn luyện được mô hình chuyển kiểu, đầu vào sẽ cần 2 thành phần gồm ảnh mục tiêu kiểu (style target image, gọi là ảnh 1) và ảnh mục tiêu nội dung (content target image, gọi là ảnh 2), đầu ra sẽ là một bức ảnh kết hợp được kiểu của anh 1 và nội dung của ảnh 2.

Hình 3-6: Từ trái sang phải là hình ảnh thật và giả lấy từ NUAA để làm ảnh kiểu mục tiêu cho thuật toán [55]

Do bài toán phát hiện giả mạo khuôn mặt có 2 lớp gồm khuôn mặt thật và khuôn mặt giả mạo nên sẽ cần 2 mô hình chuyển kiểu cho các lớp này. Ở thử nghiệm này, luận văn chọn ra 1 đối tượng trong tập dữ liệu NUAA và lấy ra 2 bức ảnh, một bức ảnh thật và

một bức ảnh giả mạo của người này (Hình 3-6) để làm ảnh mục tiêu kiểu cho 2 mô hình chuyển kiểu (thật và giả mạo) huấn luyện với tập dữ liệu OULU, lúc này tập OULU đóng vai trò như là tập các ảnh mục tiêu nội dung. Cụ thể, tập OULU được mô tả ở mục 3.1.1 được chia làm 2 tập con: Một tập gồm các ảnh thật được dùng huấn luyện cho mô hình chuyển kiểu ảnh thật, tập còn lại gồm các ảnh giả mạo được dùng huấn luyện cho mô hình chuyển kiểu ảnh giả mạo. Quá trình phân chia dữ liệu được thể hiện tại hình 3-7.



Hình 3-7: Quá trình phân chia dữ liệu cho huấn luyện mô hình chuyển kiểu

Để huấn luyện cho mô hình chuyển kiểu luận văn sử dụng các thông số đầu vào như sau:

- Số lượng epoch: 300
- Batch size: 4
- Content weight: 1e5
- Style weight: 1e10
- Learning rate: 1e-3
- Opimizer: Adam

Bởi phương pháp [55] có mục tiêu là cực tiểu hóa hàm mất mát tối ưu nhất có thể nên luận văn sẽ đánh giá phương pháp này theo hai tiêu chí định tính và định lượng. Tiêu chí định tính được thực hiện bằng cách luận văn sử dụng mô hình mới được huấn luyện để chuyển kiểu cho tập dữ liệu OULU và quan sát bằng mắt thường, trong khi đó tiêu chí định

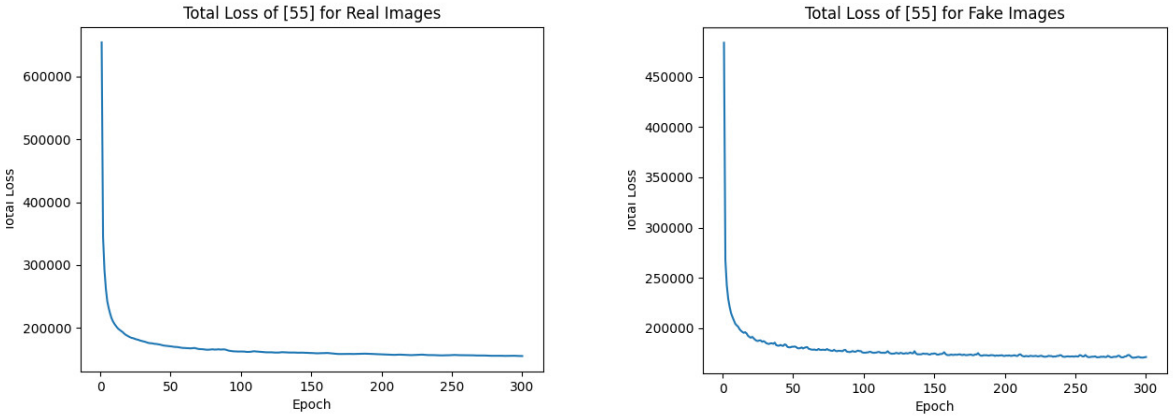
lượng sẽ được đánh giá bằng cách quan sát giá trị tổng mất mát trong suốt quá trình huấn luyện.

Về kết quả định tính, từ hình 3-8, luận văn thấy rằng dữ liệu ảnh được chuyển kiểu đang khác khá nhiều so với một hình ảnh chụp thông thường từ tập NUAA (Hình 3-6)

.

Hình 3-8: Từ trái qua phải là ảnh thật và ảnh giả mạo từ tập OULU đã chuyển kiểu

Về định lượng, hình 3-9 cho ta thấy được tổng giá trị mất mát trong toàn bộ quá trình huấn luyện, tổng giá trị mất mát nằm trong khoảng 170000 và gần như không thay đổi quá nhiều từ epoch 150.



Hình 3-9: Biểu đồ quá trình thay đổi tổng giá trị mất mát của 2 mô hình chuyển kiểu

Từ các kết quả trên, luận văn thấy rằng việc ứng dụng [55] để tạo dữ liệu mới cho bài toán phát hiện giả mạo khuôn mặt hiện chưa đạt được như kỳ vọng của ý tưởng ban đầu.

3.5. Kết luận

Chương này đã trình bày về cách thu thập dữ liệu của các bộ dữ liệu được sử dụng trong luận văn, cùng với các độ đo dùng để đánh giá các mô hình. Tiếp theo đó là các thông số cụ thể đối với mỗi thử nghiệm. Qua đó luận văn thấy rằng, với mục đích phát hiện giả mạo khuôn mặt thì việc áp dụng CDC và thông tin chiều sâu đã giảm được tỷ lệ lỗi phát hiện sai các trường hợp giả mạo thành không giả mạo, từ đó đảm bảo được cho một hệ thống phát hiện khuôn mặt trở lên đáng tin cậy và an toàn hơn. Trong khi đó, phương pháp [55] hiện tại chưa thực sự phù hợp để giải quyết vấn đề thích ứng miền trong bài toán phát hiện giả mạo khuôn mặt.

KẾT LUẬN

Luận văn tập trung nghiên cứu các phương pháp nhằm phát hiện được các khuôn mặt giả mạo trong các hệ thống phát hiện khuôn mặt và đạt được các kết quả sau:

- Thực hiện khảo sát các phương pháp phát hiện giả mạo khuôn mặt
- Nghiên cứu và áp dụng CDC, sử dụng ảnh chiều sâu của khuôn mặt vào mạng resnet-34, so sánh về tác động của các kỹ thuật này trong việc phát hiện giả mạo khuôn mặt.
- Thử nghiệm sử dụng mô hình chuyển kiểu trong vấn đề thích ứng miền

Trong tương lai, luận văn có thể tiếp tục được nghiên cứu theo hướng giảm độ lỗi trong quá trình phát hiện khuôn mặt giả mạo, đồng thời nghiên cứu theo các phương pháp khắc phục vấn đề thích ứng miền trong bài toán này.

TÀI LIỆU THAM KHẢO

[1] Z. Boulkenafet, J. Komulainen and A. Hadid (2016), "Face spoofing detection using colour texture analysis," IEEE Trans on Information Forensics and Security, vol. 11, no. 8, pp. 1818–1830.

[2] Xiaobai Li, J. Komulainen and Guoying Zhao (2016), "Generalized face anti-spoofing by detecting pulse from face videos," in Proc. of IEEE23rd Int. Conf. on Pattern Recognition, Piscataway, NJ: IEEE Press, pp. 4239–4244.

[3] S. Q. Liu, X. Y. Lan and P. C. Yuen (2018), "Remote photoplethysmography correspondence feature for 3D mask face presentation attack detection," in Proc. of the European Conf. on Computer Vision, Cham: Springer, pp. 558–573.

[4] J. K. Huang (2018), "Research on living detection technology of face recognition," Wuhan: Central China Normal University.

[5] J. Määttä, A. Hadid and M. Pietikäinen (2011), "Face spoofifing detection from single images using micro-texture analysis," in Pro. of Int. Joint Conf. on Biometrics, Colorado State, USA: IEEE, pp. 1–7.

[6] Z. Boulkenafet, J. Komulainen and A. Hadid (2017), "Face antispoofifing using speeded-up robust features and fifisher vector encoding," IEEE Signal Processing Letters, vol. 24, no. 2, pp. 141–145.

[7] C. Y. Xiang (2017), "Research on feature extraction and classification method of RGB-D images," Xiangtan: Xiangtan University.

[8] B. Peixoto, C. Michelassi and A. Rocha (2011), "Face liveness detection under bad illumination conditions," in Proc. of 18th IEEE Int. Conf. on Image Processing, Melbourne, Australia: IEEE, pp. 3557–3560.

[9] L. B. Zhang, F. Peng and L. Qin (2018), "Face spoofing detection based on color texture markov feature and support vector machine recursive feature elimination," Journal of Visual Communication and Image Representation, no. 51, pp. 56–69.

[10] R. J. Wang, J. L. Li, H. Ni, Y. J. Wu and F. Y. Huang (2015), "A face recognition method and system," China.

[11] A. K. Singh, P. Joshi and G. C. Nandi (2014), “Face recognition with liveness detection using eye and mouth movement,” in Proc. of Int. Conf. on Signal Propagation and Computer Technology, Piscataway, NJ: IEEE Press, pp. 592–597.

[12] E. S. Ng and Y. S. Chia (2012), “Face verification using temporal affective cues,” in Proc. of the 21st Int. Conf. on Pattern Recognition, Tsukuba Science City, Japan: IEEE, pp. 1249–1252.

[13] W. Bao, H. Li, N. Li and W. Jiang (2009), “A liveness detection method for face recognition based on optical flow Fifield,” in Proc. of Int. Conf. on Image Analysis and Signal Processing, Cairo, Egypt pages: IEEE, pp. 233–236.

[14] J. Galbally, S. Marcel and J. Fierrez (2014), “Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition,” IEEE Transactions on Image Processing, vol. 23 no. 2, pp. 710–724.

[15] J. Galbally and S. Marcel (2014), “Face anti-spoofifing based on general image quality assessment,” in Proc. of 22nd Int. Conf. on Pattern Recognition, Stockholm, Sweden: IEEE, pp. 1173–1178.

[16] D. Wen, H. Han and A. K. Jain (2015), “Face spoof detection with image distortion analysis,” IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 746–761.

[17] H. L. Li, S. Q. Wang and A. C. Kot (2016), “Face spoofifing detection with image quality regression,” in Proc. of 6th Int. Conf. on Image Processing Theory Tools and Applications, Oulu, Finland: IEEE, pp. 1–6.

[18] D. Yi, Z. Lei, Z. W. Zhang and Li. S. Z (2014), “Face anti-spoofifing: Multispectral approach,” Handbook of Biometric Anti-Spoofifing, Berlin: Springer.

[19] X. D. Sun, L. Huang and C. P. Liu (2016), “Context based face spoofifing detection using active near-infrared image,” in Proc. of 23rd Int. Conf. on Pattern Recognition, Cancun, Mexico: IEEE, pp. 4262–4267.

[20] X. D. Sun, L. Huang and C. P. Liu (2018), “Multispectral face spoofifing detection using VIS–NIR imaging correlation,” International Journal of Wavelets, Multiresolution and Information Processing, vol. 16, no. 2, 1840003.

[21] H. Steiner, A. Kolb and N. Jung (2016), “Reliable face anti-spoofifing using multispectral swir imaging,” in Proc. of International Conf. on Biometrics, Halmstad, Sweden: IEEE, pp. 1–8.

[22] T. Wang, J. W. Yang and Z. Lei (2013), “Face liveness detection using 3D structure recovered from a single camera,” in Proc. of Int. Conf. on Biometrics, Piscataway, NJ: IEEE Press, pp. 1–6.

[23] J. W. Yang, Z. Lei and S. Z. Li (2014), “Learn convolutional neural network for face anti-spoofing,” arXiv preprint arXiv:1408. 5601.

[24] O. Lucena, A. Junior and V. Moia (2017), “Transfer learning using convolutional neural networks for face anti-spoofing,” in Int. Conf. Image Analysis and Recognition, Springer, Cham, pp. 27–34.

[25] K. Simonyan and A. Zisserman (2014), “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556.

[26] X. Tu and Y. Fang (2017), “Ultra-deep neural network for face anti-spoofing,” in Int. Conf. on Neural Information Processing, Springer, Cham, pp. 686–695.

[27] K. He, X. Zhang and S. Ren (2016), “Deep residual learning for image recognition,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 770–778.

[28] Y. Atoum, Y. Liu and A. Jourabloo (2017), “Face anti-spoofing using patch and depth-based CNNs,” in 2017 IEEE Int. Joint Conf. on Biometrics, IEEE, pp. 319–328.

[29] Y. Liu, A. Jourabloo and X. Liu (2018), “Learning deep models for face anti-spoofing: Binary or auxiliary supervision,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 389–398.

[30] Z. Wang, C. Zhao and Y. Qin (2018), “Exploiting temporal and depth information for multi-frame face anti-spoofing,” arXiv preprint arXiv:1811.05118.

[31] X. Yang, W. Luo and L. Bao (2019), “Face anti-spoofing: Model matters, so does data,” in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, pp. 3507–3516.

[32] X. Tu, J. Zhao and M. Xie (2019), “Learning generalizable and identity-discriminative representations for face anti-spoofing,” arXiv preprint arXiv:1901.05602.

[33] K. Hornik, M. B. Stinchcombe, and H. White (1989). “Multilayer feedforward networks are universal approximators”. Neural Networks, 2(5):359–366.

[34]. Yu, R., Saito, S., Li, H., Ceylan, D., Li, H (2017) “Learning dense facial correspondences in unconstrained images”.

[35]. Bas, A., Huber, P., Smith, W.A.P., Awais, M., Kittler, J (2017) “3d morphable models as spatial transformer networks”. In: ICCV 2017 Workshop on Geometry Meets Deep Learning.

[36]. Deng, J., Cheng, S., Xue, N., Zhou, Y., Zafeiriou, S (2017) “Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition”. arXiv preprint arXiv:1712.04695.

[37]. Moschoglou, S., Ververas, E., Panagakis, Y., Nicolaou, M., Zafeiriou, S (2017) “Multiattribute robust component analysis for facial uv maps”. arXiv preprint arXiv:1712.05799

[38]. Xue, N., Deng, J., Cheng, S., Panagakis, Y., Zafeiriou, S (2018) “Side information for face completion: a robust pca approach” arXiv preprint arXiv:1801.07580

[39]. Maninchedda, F., H¨ane, C., Oswald, M.R., Pollefeys, M (2016) “Face reconstruction on mobile devices using a height map shape model and fast regularization”. In: 3D Vision (3DV), 2016 Fourth International Conference on, IEEE 489–498

[40]. Maninchedda, F., Oswald, M.R., Pollefeys, M (2017) “Fast 3d reconstruction of faces with glasses”. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE 4608–4617

[41]. Booth, J., Zafeiriou, S (2014) “Optimal uv spaces for facial morphable model construction”. In: Image Processing (ICIP), IEEE International Conference on, IEEE (2014) 4672–4676

[42]. He, K., Zhang, X., Ren, S., Sun, J (2016) “Deep residual learning for image recognition”. In: Computer Vision and Pattern Recognition. 770–778

[43]. Crispell, D., Bazik, M (2017) “Pix2face: Direct 3d face model estimation”.

[44] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, Guoying Zhao (2020), “Searching Central Difference Convolutional Networks for Face Anti-Spoofing”.

[45] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, and Zhen Lei (2018) “Exploiting temporal and depth information for multi-frame face anti-spoofing”. arXiv preprint arXiv:1811.05118. 2, 4, 6, 7, 8

[46] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, Yi Yang (2017). “Random Erasing Data Augmentation”

[47] https://towardsdatascience.com/deep-domain-adaptation-in-computer-vision-8da398d3167f. Truy cập này 15/11/2021

[48] pytorch.org. Truy cập ngày 15/11/2021

[49] Artem Rozantsev, Mathieu Salzmann, Pascal Fua (2016). “Beyond Sharing Weights for Deep Domain Adaptation”.

[50] Baochen Sun, Kate Saenko (2016). “Deep CORAL: Correlation Alignment for Deep Domain Adaptation”.

[51] Guoliang Kang, Lu Jiang, Yi Yang, Alexander G Hauptmann (2019). “Contrastive Adaptation Network for Unsupervised Domain Adaptation”.

[52] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, Nicolas Courty (2018). “DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation”.

[53] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio (2014). “Generative Adversarial Networks”

[54] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, Wen Li (2016). “Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation”.

[55] Justin Johnson, Alexandre Alahi, Li Fei-Fei (2016). “Perceptual Losses for Real-Time Style Transfer and Super-Resolution” arXiv preprint arXiv:1512.03385 (2015)

[56] Gross, S., Wilber, M (2016) “Training and investigating residual nets”. http://torch.

ch/blog/2016/02/04/resnets.html

[57] Ioffe, S., Szegedy, C (2015) “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: Proceedings of The 32nd International Conference on Machine Learning. 448–456

[58] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L (2015) “ImageNet Large Scale Visual Recognition Challenge”. International Journal of Computer Vision (IJCV) 115(3)211–252

[59] Simonyan, K., Zisserman, A (2014) “Very deep convolutional networks for large-scale image recognition”. arXiv preprint arXiv:1409.1556

BẢN CAM ĐOAN

Tôi cam đoan đã thực hiện việc kiểm tra mức độ tương đồng nội dung luận văn/luận án qua phần mềm DoIT một cách trung thực và đạt kết quả mức độ tương đồng 2% toàn bộ nội dung luận văn/luận án. Bản luận văn/luận án kiểm tra qua phần mềm là bản cứng luận văn/luận án đã nộp để bảo vệ trước hội đồng. Nếu sai tôi xin chịu hình thức kỷ luật theo quy định hiện hành của Học viện.

Hà Nội, ngày 12 tháng 09 năm 2022

HỌC VIÊN CAO HỌC/NCS

KiểmTraTàiLiệu

BÁO CÁO KIỂM TRA TRÙNG LẶP

Thông tin tài liệu

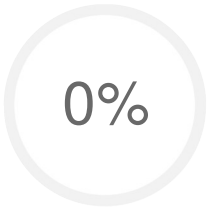
Tên tài liệu:	Luận văn Thạc Sĩ - Phạm Thanh Hùng - Final
Tác giả:	Phạm Thanh Hùng
Điểm trùng lặp:	2
Thời gian tải lên:	10:03 12/09/2022
Thời gian sinh báo cáo:	10:09 12/09/2022
Các trang kiểm tra:	78/78 trang

Kết quả kiểm tra trùng lặp

Có 2% nội dung trùng lặp

Có 98% nội dung không trùng lặp

Có 0% nội dung người dùng loại trừ



Có 0% nội dung hệ thống bỏ qua

Nguồn trùng lặp tiêu biểu

123docz.net tailieu.vn