PHƯƠNG PHÁP HỌC TƯƠNG PHẢN VÀ TỔNG HỢP ĐẶC TRƯNG CHO BÀI TOÁN CHỐNG GIẢ MẠO KHUÔN MẶT

Bùi Quốc Bảo[*], **Trần Anh Đạt**[†], **Nguyễn Khánh Hưng**[*], **Vũ Hoài Nam**[‡], **Vũ Văn Thương**[‡], **Nguyễn Việt Hưng**[‡]
[*]Đại học Bách khoa Hà Nội
[†]Đại học Thủy lợi
[‡]Học Viện Công Nghệ Bưu Chính Viễn Thông

***Abstract*—Chống giả mạo khuôn mặt (Face Anti-Spoofing, viết tắt là FAS) là một phương thức quan trọng trong các hệ thống nhận dạng khuôn mặt giúp bảo vệ và nhận dạng đúng người. Những năm gần đây, các thuật toán phát hiện giả mạo khuôn mặt được phát triển mạnh mẽ ngay cả trong các trường hợp chưa đưa vào huấn luyện mô hình thuật toán. Tuy nhiên, các mô hình thuật toán học sâu này còn khá cơ bản nên trong nhiều trường hợp mô hình vẫn chưa phát hiện được sự giả mạo khuôn mặt. Gần đây, một số mô hình đã học tập dựa trên các tín hiệu pixel để xử lý nhiệm vụ FAS. Vì vậy, trong bài báo này, nhóm tác giả đưa ra một phương pháp đánh giá mới với tên gọi FACL (Feature Aggregation and Constrastive Learning) dựa trên các đặc trưng thông tin của ảnh khuôn mặt đầu vào và học tương phản. Các thử nghiệm được xây dựng trên hai bộ dữ liệu: (1) Bộ dữ liệu của PTIT; và (2) Bộ dữ liệu của Zalo mang lại các kết quả tốt và hiệu quả hơn một số phương pháp hiện có. Ngoài ra, các nghiên cứu của tác giả cũng chứng minh mang tính hiệu quả khi mô hình học tập các lớp pixel khác nhau và đồng thời để cung cấp các thông tin chuyên sâu giúp giám sát việc chống giả mạo khuân mặt.**

***Index Terms*—face anti-spoofing, liveness detection, deep learning.**

I. GIỚI THIỆU

Trong thời đại hiện nay, việc sử dụng công nghệ nhận dạng khuôn mặt đã trở nên ngày càng phổ biến. Tuy nhiên, điều này đã mở ra cánh cửa cho các hình thức vi phạm gian lận thông qua giả mạo khuôn mặt với mức độ tinh vi ngày càng cao. Do đó, việc phát triển hệ thống chống giả mạo khuôn mặt trở nên vô cùng cấp bách và có ý nghĩa quan trọng hơn bao giờ hết. Các hệ thống này phải có khả năng xử lý các tình huống phức tạp, bao gồm việc nhận diện ảnh chụp từ các thiết bị khác, ảnh in, video được phát lại, cũng như việc phát hiện mặt nạ 3D và các phương pháp giả mạo khác.

Công nghệ nhận dạng khuôn mặt hiện nay đã đạt được sự tiến bộ đáng kể với độ chính xác cao [1]. Điều này được thúc đẩy bởi sự phát triển đáng kể của các bộ dữ liệu, trong đó nhiều nhóm nghiên cứu [2], [3] đã thu thập một lượng lớn thông tin về khuôn mặt con người từ khắp nơi trên thế giới. Sự gia tăng đột biến này về khối lượng dữ liệu đã cung cấp nền tảng cho việc phát triển các thuật toán học sâu trong việc nhận dạng khuôn mặt. Tuy nhiên, nhiều dữ liệu vẫn chưa trải qua quá trình xử lý chất lượng, dẫn đến tăng cường về số lượng dữ liệu nhưng chất lượng chưa được đảm bảo. Do đó, nhiều nhóm nghiên cứu [4] đã sử dụng những dữ liệu này để xây dựng các mô-đun tấn công vào các hệ thống nhận diện khuôn mặt.

Các bộ dữ liệu được sử dụng để xây dựng các mô-đun tấn công vào hệ thống nhận diện khuôn mặt thường chủ yếu bao gồm ảnh RGB [5]. Nhược điểm của các bộ dữ liệu này thường xuất phát từ việc hạn chế về số lượng các đối tượng được bao gồm. Ngoài ra, cũng có một số bộ dữ liệu [5] được sử dụng để

Tác giả liên hệ: Nguyễn Việt Hưng,
Email: nvhung_vt1@ptit.edu.vn
Đến tòa soạn: 10/2023, chỉnh sửa:11/2023, chấp nhận đăng: 12/2023.

phát triển các mô hình chống giả mạo. Những bộ dữ liệu này được biết đến với sự đa dạng về kích thước và định dạng, bao gồm cả ảnh RGB, hồng ngoại (IR) và thông tin độ sâu.

Trong nghiên cứu này, nhóm tác giả giới thiệu một phương pháp mới nhằm giải quyết thách thức về việc ngăn chặn giả mạo thông qua nhận diện khuôn mặt. Nhóm tác giả đã thay đổi cấu trúc của mô hình, tiến hành xử lý từng phương pháp riêng biệt và kết hợp các đặc trưng từ lớp pixel ở các cấp độ khác nhau, nhằm tăng cường sự tương tác giữa các nhánh thông tin RGB, hồng ngoại (IR), và độ sâu trong mạng nơ-ron. Thiết kế này cho phép tổng hợp các đặc trưng thông tin của ảnh khuôn mặt đầu vào (*Feature Aggregation*) và kết hợp với phương pháp học tương phản (*Contrastive Learning*), viết tắt là FACL. Trong các phần tiếp theo, chúng ta sẽ đi sâu vào tìm hiểu chi tiết phương pháp đề xuất này.

II. NGHIÊN CỨU LIÊN QUAN

Có hai cách tiếp cận để phát hiện giả mạo khuôn mặt là: (1) Các Phương pháp Truyền thống; và (2) Các Phương pháp Học sâu.

A. Các Phương pháp Truyền thống:

Sự khác biệt về cấu trúc là một trong những dấu hiệu chính để phân biệt giữa khuôn mặt thật và khuôn mặt giả mạo [6]. Thông tin như vậy đã được khai thác cho việc chống giả mạo khuôn mặt. Ví dụ, nhiều đặc trưng được tạo thủ công đã được nghiên cứu trong các công trình trước đó, bao gồm LBP [7] Trích xuất đặc trưng về cấu trúc bằng cách so sánh mỗi pixel với các pixel xung quanh, tuy nhiên nhược điểm của phương pháp này với sự biến đổi về ánh sáng và hướng, HOG [8] Mô tả sự phân phối của độ dốc hoặc hướng biên qua các khối chồng chéo, thích hợp cho việc phát hiện đối tượng nhưng tạo ra các vectơ đặc trưng có số chiều lớn., DOG [9] Phát hiện cực trị về độ sáng ở nhiều tỷ lệ khác nhau, không thay đổi theo hướng, nhưng có thể hoạt động không tốt cho cấu trúc không giống như vết đen., SIFT [10] Trích xuất các tính năng bất biến có tính đặc biệt cao từ ảnh, mạnh mẽ trước nhiều biến thể nhưng tốn chi phí tính toán. và SURF [11] lấy cảm hứng từ SIFT nhưng tính toán nhanh hơn. Ngoài ra, các miền dữ liệu khác nhau đã được khai thác để trích xuất các đặc trưng phân biệt. Boulkenafet và cộng sự đã điều tra các không gian màu khác nhau như HSV và YCbCr [12] cung cấp thông tin bổ sung so với ảnh đa cấp. HSV tách độ sáng khỏi thông tin màu sắc trong khi YCbCr mô hình hóa quan sát thị giác của con người. Các đặc trưng trong miền tần số cũng được nghiên cứu trong [8]. Chúng hiệu quả trong phân loại cấu trúc vật liệu nhưng nhạy cảm với các dịch chuyển. Vấn đề phổ biến tồn tại trong những phương pháp này là các đặc trưng được tạo thủ công không mạnh mẽ trước các biến số phiền hà khác nhau trong môi trường thực tế, như ánh sáng và che khuất.

Khác với việc chỉ sử dụng hình ảnh tĩnh, các nhà nghiên cứu cố gắng tận dụng các chuyển động khuôn mặt tự nhiên trong một chuỗi khung hình cho việc chống giả mạo khuôn mặt. Ví dụ, việc chớp mắt đã được sử dụng để phát hiện sự sống động của khuôn mặt trong [13]. Chavan và cộng sự đã sử dụng các chuyển động của miệng và môi cho việc chống giả mạo khuôn mặt [14]. Tuy nhiên, các chuyển động khuôn mặt tự nhiên thường quá tinh tế để được ghi lại bằng các đặc trưng tạo thủ công trong thực tế.

B. Các Phương pháp Học sâu:

Sức mạnh biểu diễn mạnh mẽ của các mạng CNN hiện đại đã được khai thác trong nghiên cứu chống giả mạo khuôn mặt [15]. Các phương pháp trong [16] đã sử dụng một mô hình CaffeNet hoặc VGG-face được huấn luyện trước như một bộ trích xuất đặc trưng để phân biệt giữa khuôn mặt thật và giả mạo. Nhưng nó cũng đối mặt với thách thực dễ bị tấn công bằng cách áp dụng các biện pháp thêm nhiễu vào ảnh, kỹ thuật này có thể làm suy giảm khả năng phân biệt của mô hình và tạo ra kết quả giả mạo. Mô hình LSTM-CNN [17] với khả năng vượt trội trong xử lý chuỗi dữ liệu và giữ lại thông tin trạng thái trước đó để nắm bắt mối quan hệ thời gian và chuỗi trong dữ liệu hình ảnh nhưng đòi hỏi lượng tài nguyên tính toán đáng kể và có thể không linh hoạt đối với dữ liệu phức tạp. Hay với kiến trúc mạng GAN [23] có thể hỗ trợ tạo ra dữ liệu giả mạo để nâng cao hiệu suất đào tạo mô hình. Tuy nhiên, đối mặt với nhược điểm lớn khi đối phó với các kỹ thuật giả mạo tiên tiến như deepfake, có thể đặt thách thức đối với độ tin cậy của hệ thống phát hiện.

C. Phương pháp Tổng quát hóa miền

Những nghiên cứu gần đây về Tổng quát hóa Miền (domain generalization) trong lĩnh vực chống giả mạo khuôn mặt với mục tiêu là xây dựng các mô hình có khả năng phát hiện tốt ngay cả trong những tình huống đa dạng và không được biết trước. Ví dụ, bài

báo mới [21] đã đề xuất sử dụng kỹ thuật tạo dữ liệu tiêu cực (negative) thay vì sử dụng các mẫu tấn công thực tế trong quá trình đào tạo. Một hướng tiếp cận khác [22] sử dụng mô hình dựa trên năng lượng (energy-based model, EBM), với mục tiêu khuyến khích các ảnh khuôn mặt thật có giá trị hàm năng lượng tự do thấp và xem xét tất cả các mẫu có năng lượng cao là các khuôn mặt giả mạo. Bên cạnh đó, các phương pháp khác [19], [20] cũng đã mở ra những hướng tiếp cận mới cho vấn đề tổng quát hóa miền.

Mặc dù đã có sự tiến bộ đáng kể, việc đảm bảo tính tổng quát hóa trong bài toán chống giả mạo khuôn mặt khi chuyển đổi giữa các miền dữ liệu vẫn là một thách thức lớn. Các nghiên cứu trong lĩnh vực này đang liên tục nỗ lực để nâng cao khả năng ứng dụng thực tế của các hệ thống chống giả mạo khuôn mặt.

III. PHƯƠNG PHÁP ĐỀ XUẤT

A. Bài toán chống giả mạo khuôn mặt tổng quát hóa miền

Bài toán chống giả mạo khuôn mặt là một trong những ứng dụng quan trọng của công nghệ nhận dạng khuôn mặt và bảo mật dữ liệu. Mục tiêu của bài toán này là ngăn chặn và phát hiện các hoạt động giả mạo, làm nhái hoặc sử dụng sai trái thông tin về khuôn mặt. Cụ thể, khi sử dụng các ứng dụng liên quan, thay vì sử dụng ảnh chụp khuôn mặt trực tiếp, người dùng có thể dùng ảnh của một người khác được chụp gián tiếp qua các thiết bị điện tử để giả mạo danh tính. Nhiệm vụ của chúng ta là xây dựng một mô hình học máy có khả năng phân biệt được các trường hợp ảnh thật hay ảnh giả mạo. Về mặt bản chất, đây là một bài toán phân loại nhị phân với ảnh giả mạo (nhãn là 0) và ảnh thật (nhãn là 1). Tuy nhiên, các mô hình phân loại thông thường hoạt động không hiệu quả với các phương thức giả mạo mới và không có tính tổng quát hóa tốt. Do đó, nhóm tác giả đề xuất một phương pháp dựa trên khái niệm tổng quát hóa đa miền (cross-domain generalization) để tăng khả năng tổng quát hóa của mô hình học máy.

Đối với phương pháp này, trong dữ liệu huấn luyện, lớp các ảnh giả mạo sẽ được chia thành các phân lớp con ứng với các miền (domain) khác nhau thể hiện các phương thức giả mạo khác nhau. Cụ thể, cho tập huấn luyện với N phần tử $D = \{(x_i, y_i, e_i)\}(i = 1)^N$, trong đó $x_i \in X$, $y_i \in Y$ lần lượt là ảnh đầu vào và nhãn tương ứng. Với không gian ảnh RGB X, không gian đầu ra $Y = \{0(live), 1(spoof)\}$ và E miền $E = \{e^{(1)}, e^{(2)}, .., e^{(E)}\}$, mục tiêu là xác định một hàm quyết định f:

$$f : x \rightarrow \{0(live), 1(spoof)\} \quad (1)$$

phân loại xem một mẫu x từ miền dữ liệu mới $e^{'}$ có phải là thật (live) hay giả mạo (spoof).

B. Kiến trúc mô hình

Trong bài báo này, nhóm tác giả đề xuất sử dụng kỹ thuật trích xuất đặc trưng chứa thông tin về phong cách (Style Information) và đặc trưng thông tin về nội dung (Content Information) để tách biệt các đặc tính toàn cục và địa phương của ảnh trong bài toán FAS. Đồng thời, phương pháp chuyển đổi phong cách (style transfer) được áp dụng giống như một cách để tăng cường và đa dạng hóa các đặc trưng giúp cải thiện khả năng tổng quát hoá của mô hình. Những kỹ thuật này được đề xuất lần đầu trong nghiên cứu của Zhuo Wang và các cộng sự [19], cho các kết quả tốt trên các tập dữ liệu chống giả mạo nổi tiếng đã được công bố. Dựa trên các ý tưởng nền tảng đó, trong phần này, tác giả sẽ trình bày kiến trúc mạng tổng quan của mình với bốn module mạng cụ thể như dưới đây.

1) Module xương sống trích xuất các đặc trưng cơ bản: Đầu tiên, ảnh đầu vào ban đầu x sẽ được xử lý qua một mạng xương sống (backbone) để trích xuất các đặc trưng cơ bản. Các mô hình backbone có thể kể đến như VGG, ResNet, EfficientNet, ... Dựa trên đánh giá về thời gian xử lý và độ chính xác của các mạng trên các tập dữ liệu nổi tiếng đã được công bố (như ImageNet, CIFAR-100, v.v), trong các thực nghiệm ở phần IV của bài báo, tác giả lựa chọn kiến trúc ResNet-18 làm mô hình xương sống. Đồng thời, các trọng số đã được huấn luyện (pre-trained weights) cũng được sử dụng để tăng khả năng của toàn bộ mô hình.

Mạng ResNet sử dụng các khối cơ bản gọi là "residual blocks" để xây dựng cấu trúc mạng. Trong kiến trúc của ResNet-18, tác giả sử dụng bốn đặc trưng $x_{(1\_1)}, x_{(1\_2)}, x_{(1\_3)}, x_{(1\_4)}$ từ bốn khối cuối với thông tin chiều lần lượt là (64, 64, 64), (128, 32, 32), (256, 16, 16) và (512, 8, 8) (ba chiều của đặc trưng lần lượt là chiều sâu hay số lượng kênh, chiều cao và chiều rộng). Bốn đặc trưng này sau đó sẽ được đi qua hai module trích xuất thông tin về nội dung và phong cách.

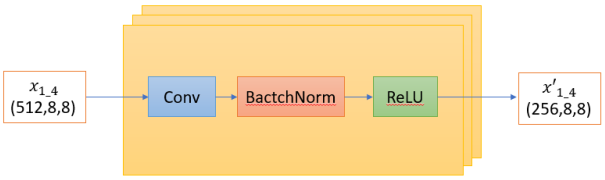

Figure 1. Kiến trúc module trích xuất thông tin về nội dung.

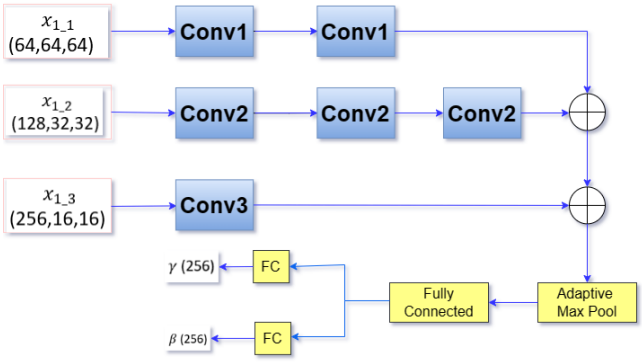

Figure 2. Kiến trúc module trích xuất thông tin về phong cách.

2) Module trích xuất thông tin về nội dung: Trong hệ thống nhận diện khuôn mặt, thông tin về nội dung thường được biểu diễn bằng các đặc điểm thông thường, bao gồm cả đặc điểm ngữ nghĩa và thuộc tính vật lý của hình ảnh. Với hình ảnh khuôn mặt, thường có sự chia sẻ của không gian đặc trưng ngữ nghĩa chung. Thậm chí, bất kể hình ảnh có phải là thật hay giả, các đặc điểm về hình dáng và kích thước thường khá giống nhau trong mỗi bức ảnh. Để trích xuất thông tin về nội dung, chúng ta sử dụng đặc trưng từ block cuối cùng của cấu trúc chung. Đặc trưng này có khả năng mô tả chi tiết và mang thông tin tổng quan hơn so với các khối trước đó.

Kiến trúc của model bao gồm 3 lớp tích chập, BatchNorm và ReLU được xếp chồng lên nhau để tạo thành một mạng sâu. Lớp tích chập để tạo ra một biểu diễn đặc trưng của hình ảnh. Các lớp tích chập sử dụng các kernel có kích thước nhỏ để quét qua hình ảnh và trích xuất thông tin về cấu trúc và đặc điểm quan trọng. Lớp Batch Normalization được sử dụng để đảm bảo rằng phân phối đầu ra của lớp tích chập ổn định, giúp tăng tốc quá trình đào tạo và cải thiện hiệu suất. Hàm kích hoạt ReLU được áp dụng sau mỗi lớp tích chập để tạo tính phi tuyến tính và kích thích việc học các đặc trưng phức tạp trong dữ liệu. Kết quả cuối cùng ta thu được đặc trưng nội dung $x^{'}_{(1\_4)}$ với chiều là $(256, 8, 8)$, chứa thông tin quan trọng về nội dung hình ảnh.

3) Module trích xuất thông tin về phong cách: Thông tin về phong cách (Style Information Feature) thường liên quan đến cách mà các yếu tố như biến đổi hình dáng, màu sắc, và cấu trúc tạo nên một phong cách riêng biệt cho một hình ảnh cụ thể. Đây là một khái niệm được giới thiệu và đề xuất trong một bài báo của một nhóm tác giả [10]. Khi chúng ta sử dụng GradCAM để trực quan hóa các đặc trưng của hình ảnh, chúng ta thấy rằng các đặc trưng về nội dung thường tập trung vào phần chứa khuôn mặt, trong khi đặc trưng về phong cách tập trung chủ yếu vào phần bao quanh và nền của ảnh. Trong các mạng nơ-ron tích chập, hình ảnh được biểu diễn dưới dạng các "đặc trưng" ở các lớp khác nhau của mạng. Đặc trưng thông tin về phong cách là một phần của những đặc trưng này và chứa thông tin về các mẫu, biến đổi hình dáng và các đặc điểm không gian của ảnh. Do đó, để thu thập thông tin về phong cách, chúng ta sử dụng một phương pháp dựa trên mô hình kim tự tháp, thu thập các đặc trưng đa tầng cùng với cấu trúc phân cấp.

Trong phần khối ban đầu của mạng, mục tiêu là tạo ra các đặc trưng phù hợp với đặc trưng x_{1_2} khi đưa qua một lớp Convolution bằng cách đưa đặc trưng x_{1_1} qua hai lớp Convolution. Các lớp Convolution này giúp biến đổi x_{1_1} thành một đặc trưng mới với các chiều tương tự như x_{1_2}. Sau đó, hai đặc trưng này được kết hợp lại với nhau thông qua một phép kết hợp đặc trưng, mục đích là cung cấp khả năng biểu diễn đa dạng hơn về thông tin trong ảnh. Tương tự, trong khối thứ hai, đặc trưng x_{1_3} trải qua một lớp Convolution để tạo ra một đặc trưng mới. Sau đó, đặc trưng mới này được kết hợp với đặc trưng kết quả từ x_{1_1} và x_{1_2} thông qua phép kết hợp đặc trưng, mục đích là giúp đa dạng hóa thông tin trong ảnh. Khối cuối cùng của mạng đảm bảo rằng kích thước đầu ra giữ nguyên kích thước đầu vào. Điều này đảm bảo rằng thông tin từ các khối trước đó không bị mất. Các tham số của lớp Convolution trong khối cuối cùng có thể được điều chỉnh tùy theo yêu cầu cụ thể của bài toán. Sau khi đi qua khối cuối cùng, vector đặc trưng thu được được đưa vào một toán tử Adaptive Max Pooling để giảm kích thước của nó. Cuối cùng, vector này sẽ trải qua các lớp Fully Connected để trích xuất hai vector tham số affine, mỗi vector có số chiều là 256. Các lớp Fully Connected có thể được tùy chỉnh về số lượng, kích thước và hàm kích hoạt tùy thuộc vào nhiệm vụ cụ thể của bài toán. Các

đặc trưng từ các lớp tích chập và kết nối đầy đủ này cung cấp thông tin phong cách đa dạng và có khả năng biểu diễn các đặc điểm của hình ảnh một cách toàn diện.

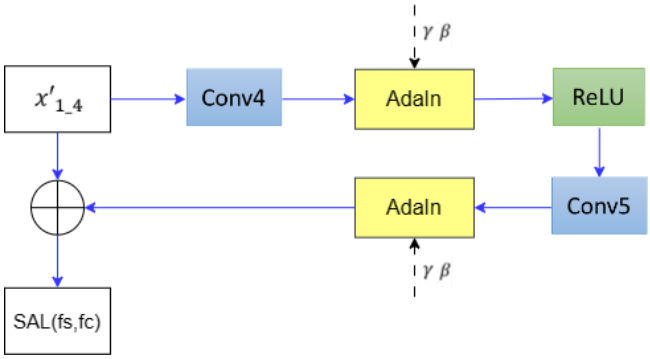

Figure 3. Kiến trúc module chuyển đổi và tổng hợp phong cách, nội dung.

4) Module chuyển đổi và tổng hợp phong cách, nội dung: Mục đích chính của module này là để tổng hợp và xáo trộn giữa đặc trưng nội dung và phong cách. Trong kiến trúc này, bên cạnh các lớp tích chập conv 4 và 5, tác giả bổ sung thêm các lớp AdaIN - được sử dụng để điều chỉnh phong cách giữa các hình ảnh. Điều này cho phép ta áp dụng phong cách từ một hình ảnh nguồn lên một hình ảnh đích mà không cần học các tham số phong cách.

Thường thì việc áp dụng phong cách từ hình ảnh nguồn lên hình ảnh đích đòi hỏi việc huấn luyện mô hình đặc biệt để thực hiện điều này. AdaIN giải quyết vấn đề này bằng cách sử dụng tích hợp của hai phần: Adaptive Normalization và Instance Normalization. Trong đó:

- Instance Normalization: Đây là một kỹ thuật chuẩn hóa dữ liệu trong mỗi bản đồ đặc trưng (feature map) riêng lẻ (instance). Nó giúp giảm biến động trong phân phối của các đặc trưng và giúp mô hình học được các tính năng chung của hình ảnh.
- Adaptive Normalization: Đây là phần động của AdaIN. Nó cho phép thay đổi phân phối của các đặc trưng trong mỗi kênh (channel) dựa trên phân phối của các đặc trưng từ hình ảnh nguồn.

Cụ thể, trong AdaIN, ta sẽ sử dụng thông tin phong cách từ hình ảnh nguồn để điều chỉnh tham số chuẩn hóa và trung bình trong Instance Normalization của hình ảnh đích. Điều này dẫn đến việc áp dụng phong cách từ hình ảnh nguồn lên hình ảnh đích một cách tùy chỉnh. Với một đặc trưng nội dung đầu vào là x và các tham số affine θ, β được trích xuất từ module phong cách, ta có công thức tổng quát của AdaIN như sau:

$$AdaIN(x, \theta, \beta) = \theta((x - \phi(x))/(\sigma(x))) + \beta, \quad (2)$$

Khái niệm về đặc trưng phong cách không chỉ bao gồm thông tin liên quan đến độ chính xác, mà còn chứa thông tin cụ thể cho từng lĩnh vực riêng biệt, điều này có thể gây ra sự thiên lệch trong quá trình tối ưu hóa mạng. Để khắc phục vấn đề này, phương pháp tổ hợp xáo trộn đặc trưng phong cách để tạo ra các đặc trưng phong cách bổ trợ, giúp cải thiện khả năng tổng quát hóa trong lĩnh vực cụ thể. Quá trình đó được biểu diễn bằng công thức sau:

$$SAL(f_c(x_i), f_s(x_i)), \quad (3)$$

trong đó f_c là đặc trưng về nội dung, f_s là đặc trưng về phong cách và x_i đại diện cho các mẫu đầu vào. Quá trình tổ hợp phong cách sẽ sử dụng cả đặc trưng nội dung và phong cách tạo ra một không gian đặc trưng ghép cặp. Quá trình xáo trộn phong cách sẽ sử dụng $f_s(x_i)$ là ngẫu nhiên tạo nên tổ hợp được xáo trộn.

Cuối cùng, ta sẽ thu được một đặc trưng tổng hợp đa dạng và tối ưu với số chiều là 256.

5) Hàm mất mát: Sau khi mô tả cụ thể các phương thức hoạt động của mạng ở phần trên, trong phần này, nhóm tác giả tổng hợp các hàm mất mát bao gồm hàm Cross Entropy và hàm tương phản (Contrastive Loss) để xây dựng một hàm mất mát tổng cho quá trình huấn luyện mô hình một cách ổn định và tối ưu.

Hàm mất mát đối chứng (contrastive loss) được áp dụng để đối phó với vấn đề liên quan đến các đặc trưng phong cách trong quá trình tối ưu hóa mạng, đặc biệt là trong bối cảnh chuyển đổi miền. Vấn đề cơ bản là các đặc trưng phong cách dựa theo lĩnh vực cụ thể có thể che khuất hoặc làm mất đi các đặc trưng quan trọng liên quan đến tính sống còn. Để khắc phục vấn đề này, một phương pháp học đối chứng được đề xuất nhằm làm nổi bật các đặc trưng phong cách liên quan đến tính sống còn và đồng thời tạo sự kìm nén đối với các đặc trưng phong cách riêng biệt cho từng miền. Sau khi kết hợp các đặc trưng nội dung và phong cách, chúng ta thu được hai tập hợp đặc trưng: một tập được gọi là $F1$ (đặc trưng tự ghép cặp) và một tập được gọi là $F2$ (đặc trưng tổ hợp bị xáo trộn). Tập đặc trưng $F1$ được đưa vào một bộ phân loại và được giám sát bằng tín hiệu thực thể nhị

phân sử dụng hàm mất mát L_{cls}.Tập đặc trưng $F2$ được so sánh với $F1$ bằng cách sử dụng độ tương tự cosin chuẩn hóa l_2 nhằm đo lường sự khác biệt giữa chúng. Các đặc trưng tự ghép cặp trong $F1$ được coi như các điểm neo trong không gian đặc trưng đã được điều chỉnh theo phong cách. Một phép đặt dừng gradient (stop-gradient) được áp dụng trên tập $F1$ để giữ vị trí của chúng không thay đổi trong không gian đặc trưng. Các đặc trưng trong tập $F2$ sau đó được hướng dẫn để tiến gần hoặc xa khỏi các điểm neo tương ứng trong tập $F1$ dựa trên thông tin về tính xác thực. Trong quá trình này, lan truyền ngược chỉ được áp dụng qua các đặc trưng trong tập $F2$ và không được áp dụng cho các đặc trưng trong tập $F1$, đồng thời thông tin phong cách được tập trung vào tính xác thực để tạo ra đặc trưng đối chứng. Hàm mất mát đối chứng L_{contra} được tính toán dựa trên sự tương quan giữa $F1$ và $F2$ cùng với thông tin về tính sống còn trong các đặc trưng. Do đó, hàm mất mát đối chứng L_{contra} có thể được biểu thị như sau:

$$L_{contra} = \begin{cases} \sum_{i=1}^{N}(+1) \cdot Sim(stopgrad(F1), F2), \\ \sum_{i=1}^{N}(-1) \cdot Sim(stopgrad(F1), F2) \end{cases} \quad (4)$$

trong đó $+1$ và -1 đo lường tính nhất quán của các nhãn về tính xác thực giữa 2 đặc trưng đầu vào.

IV. KẾT QUẢ THỰC NGHIỆM

A. Bộ dữ liệu huấn luyện và giao thức đánh giá

Trong các hệ thống chống giả mạo trong thực tế, chúng ta thường gặp hai kiểu giả mạo chính bao gồm: ảnh chụp gián tiếp qua các ảnh vật lý và ảnh chụp gián tiếp qua thiết bị điện tử như máy tính, điện thoại, ti vi, v.v. Đối với kiểu giả mạo thứ hai chúng ta lại có thể chia thành hai lớp nhỏ, chi tiết hơn, gồm ảnh gián tiếp mà phần thiết bị gián tiếp được chụp rõ và trường hợp thứ hai là các ảnh mà ta không nhìn rõ thiết bị đó. Trong phần thực nghiệm của bài báo này, nhóm tác giả sử dụng bốn bộ dữ liệu giả mạo với tên gọi như sau: PFP, CRFP, URFP, và ZFF; một bộ dữ liệu ảnh khuôn mặt thật là RFP. Cụ thể, để áp dụng phương pháp cross-domain, tác giả coi các dữ liệu hiện tại theo bốn domain như dưới đây:

- Domain 1 (ký hiệu là P) gồm tập dữ liệu PFP, viết tắt của cụm từ Photo Face PTIT, bao gồm các ảnh chụp gián tiếp qua các ảnh vật lý.
- Domain 2 (ký hiệu là C) - tương ứng với tập dữ liệu CRFP (Clear Replay Face PTIT) bao gồm các ảnh chụp gián tiếp qua thiết bị điện tử như máy tính, điện thoại, v.v. Trong đó, phần background của các ảnh này có thể nhìn rõ phần thiết bị trong khung hình. Con người có thể dễ dàng nhận diện được các trường hợp này là giả mạo.
- Domain 3 (ký hiệu là U), gồm tập URFP (Unclear Replay Face PTIT) trong đó các ảnh chụp gián tiếp qua thiết bị điện tử đã được đã được crop chi tiết vào phần ảnh chân dung. Các dữ liệu này có thể dễ gây nhầm lẫn, dẫn đến việc khó khăn hơn trong việc nhận diện, ngay cả với con người.
- Domain 4 (ký hiệu là Z) bao gồm các dữ liệu giả mạo của Zalo, được công bố trong cuộc thi Zalo AI Challenge năm 2022, với phương thức và phong cách chụp khác so với ba bộ dữ liệu kể trên.
- Bộ dữ liệu RFP (Real Face PTIT): được chia đều thành bốn phần bằng nhau và phân bổ về bốn Domain nói trên.

Datasets	live	not_live
Domain 1	3770	698
Domain 2	3770	983
Domain 3	3771	2852
Domain 4	4748	4529

Table I
THỐNG KÊ SỐ LƯỢNG ỨNG VỚI 4 DOMAIN.

(*Lưu ý:* Các dữ liệu thuộc Domain C, U và Z về bản chất đều thuộc phương thức giả mạo chụp gián tiếp qua các thiết bị điện tử; tuy nhiên, các dữ liệu này có các phong cách khác nhau. Việc chia nhỏ thành các domain sẽ giúp phần thực nghiệm để đánh giá khả năng tổng quát hóa của mô hình.)

Thông tin cụ thể của các tập dữ liệu trên được thống kê chi tiết như trong bảng I

B. Cài đặt và cấu hình thực nghiệm

Cài đặt Loại Bỏ Một - Leave-One-Out (LOO). Đối với đánh giá tổng quan, nhóm tác giả tiến hành kiểm tra chéo các tập dữ liệu bằng cách sử dụng chiến lược LOO: ta chọn ba tập dữ liệu để huấn luyện và một tập dữ liệu còn lại để kiểm tra. Cụ thể, ta có các cấu hình thực nghiệm như sau PCU $\rightarrow$ Z là giao thức huấn luyện trên ba tập PFP, CRFP, ZFF và kiểm tra trên tập URFP. PCZ $\rightarrow$ U, PUZ $\rightarrow$ C và CUZ $\rightarrow$ P được định nghĩa theo cách tương tự.

Cấu hình huấn luyện. Mô hình đề xuất nhận input là ảnh đầu vào với kích thước được điều chỉnh

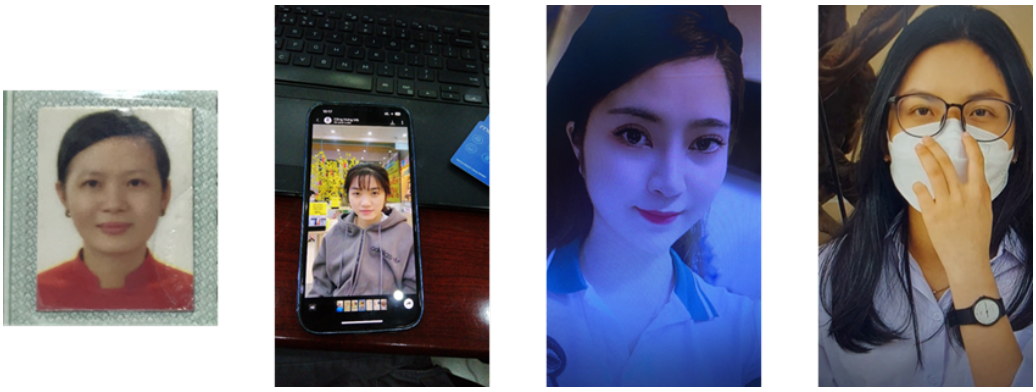

Figure 4. Kiến trúc module trích xuất thông tin về nội dung.

thành 256x256, kiến trúc mạng backbone sử dụng là ResNet-18. Tác giả huấn luyện mạng bằng bộ tối ưu hóa SGD với tốc độ học ban đầu là 5e-3, được giảm đi 2 lần tại các epoch 40 và 80, tổng số epoch huấn luyện là 100 epoch. Bên cạnh đó, kích thước batch là 96, mô hình huấn luyện trên 2 GPU (GeForce GTX 1080 Ti) với RAM 11GB trên mỗi GPU tương ứng.

Các độ đo đánh giá mô hình. Tác giả đánh giá hiệu suất mô hình bằng ba thước đo tiêu chuẩn: Tỷ lệ lỗi Half Total (HTER), diện tích dưới đường cong (AUC) và độ chính xác (ACC). Cụ thể:

- HTER là sự kết hợp của tỷ lệ False Acceptance Rate (FAR) và tỷ lệ False Rejection Rate (FRR) được tính bằng công thức sau:

$$HTER = \frac{FAR + FRR}{2}$$

  HTER đo lường sự cân bằng giữa việc chấp nhận sai và bỏ lỡ sai, với mục tiêu là tối ưu hóa sự cân bằng này. HTER càng thấp thì hiệu suất của hệ thống càng tốt.
- AUC thường được tính bằng diện tích nằm dưới đường cong ROC (Receiver Operating Characteristic). ROC được xây dựng bằng cách biểu thị tỷ lệ True Positive Rate (TPR) và tỷ lệ False Positive Rate (FPR) khi ngưỡng thay đổi từ 0 đến 1. Giá trị AUC nằm trong khoảng từ 0 đến 1, trong đó 1 đại diện cho mô hình hoàn hảo và 0.5 đại diện cho mô hình ngẫu nhiên.
- ACC đơn giản được tính bằng cách chia số lượng dự đoán đúng (True Positives và True Negatives) cho tổng số lượng mẫu dữ liệu trong tập kiểm tra.

C. So sánh các kết quả số

Nhóm tác giả lần lượt so sánh các kết quả số theo các tình huống sau đây:

- Để đánh giá khả năng tổng quát hóa đa miền của mô hình. Tác giả lần lượt so sánh các kết quả của mô hình đề xuất với một số mô hình khác theo bốn giao thức kể trên.
- Đánh giá và phân tích chi tiết các mô hình theo ma trận confusion và các chỉ số đánh giá khác.
- Đánh giá kích cỡ và hiệu quả thời gian của các mô hình qua số tham số và thời gian xử lý mỗi ảnh.

1) Đánh giá tính tổng quát hóa theo bốn giao thức dữ liệu: Để thực hiện đánh giá hiệu suất của mô hình đề xuất, nhóm tác giả so sánh kết quả với một mô hình phân loại thông thường ResNet và bốn mô hình mới nhất bao gồm D2 AM, SDA, DRDG, và SSAN.

Các mô hình trên được đánh giá chéo trên tập dữ liệu ResNet-18, SSAN, SAFAS và mô hình đề xuất được đánh giá theo bốn tình huống phân chia dữ liệu, theo bốn giao thức thực nghiệm bao gồm: PCU→Z, PCZ→U, PUZ→C và CUZ→P. Kết quả từ Bảng II cho thấy rằng trong phần lớn các trường hợp, mô hình của tác giả cho kết quả tốt hơn so với các phương pháp còn lại. Bên cạnh đó, SSAN - một kiến trúc cũng sử dụng kết hợp hai loại đặc trưng thông tin, cho kết quả ổn định, chỉ đứng sau FACL về các chỉ số. Đặc biệt, AUC cho giao thức PCU → Z cho kết quả vượt trội hơn cả. Ngược lại, trong các mô hình, hiệu suất của ResNet-18 là thấp nhất, tiếp đến SDA và D2 AM. Mô hình đề xuất cho kết quả khá ổn định nhưng chưa đáng kể và một vài trường hợp cho kết quả thấp hơn các phương pháp đã công bố có thể giải thích là do phương thức đánh giá của bài toán tổng quát hóa miền dựa trên cơ chế leave-one-out, dữ liệu đánh giá có những đặc trưng khác biệt khá nhiều so với dữ liệu huấn luyện nên cải tiến của những phương

pháp này thường không thể hiện vượt trội, mức độ cải tiến trung bình của chúng tôi là $1,2\%$ là kết quả chấp nhận được và có thể chứng tỏ độ hiểu quả của PP đề xuất. Các kết quả thử nghiệm này chỉ ra rằng mô hình được đề xuất cũng hữu ích để cung cấp hướng dẫn đa quy mô phong phú trên nhiều nguồn domain của dữ liệu.

2) Đánh giá cụ thể khả năng phân loại của mô hình: Trước tiên, chúng ta quan sát Hình 5, để theo dõi hoạt động của hàm mất mát theo từng epoch. Về mặt tổng quan, hàm dao động lên xuống khá mạnh nhưng nhìn chung theo số epoch tăng dần cũng có xu hướng giảm. Điều này phần nào cho thấy mô hình đã học được từ dữ liệu huấn luyện.

Khi triển khai mô hình chống giả mạo khuôn mặt trong thực tế, ta cần xác định một giá trị ngưỡng tối ưu để hệ thống có thể đưa ra quyết định xem ảnh khuôn mặt đầu vào có phải ảnh thật hay không. Do đó, việc sử dụng các bảng thống kê như ma trận confusion là rất cần thiết để chúng ta có thể tìm ra ngưỡng tốt nhất. Hai bảng III, IV dưới đây, thể hiện ma trận nhầm lẫn của hai mô hình tốt nhất SSAN và mô hình nhóm tác giả đề xuất FACL, với các giá trị ngưỡng tối ưu của hai phương pháp lần lượt là 0.7 và 0.6. Các bảng này trình bày các giá trị của Giá trị dương tính thực (True Positive) và Âm tính thực (True Negative) là rất lớn so với giá trị của Giá trị dương tính giả (False Positive) và Âm tính giả (False Negative), đó là lý do giải thích tại sao độ chính xác cao so với kết quả đạt được sau khi đánh giá. Dựa vào đây, nhóm tác giả nhận định rằng tỉ lệ sai của mô hình của mình là chấp nhận được trong các hệ thống xác thực chống giả mạo.

Hai bảng cho thấy ưu điểm của FACL so với SSAN khi giá trị ngưỡng nhỏ hơn trong khi các tỷ lệ nhận diện sai FP và FN nhỏ hơn.

Vẫn tương tự như các bảng kết quả trước, trong bảng V mô hình ResNet-18 cho kết quả thấp nhất và mô hình đề xuất cho kết quả cao như 91,88% với độ đo AUC và độ chính xác là 92%. Nhìn chung, với 2 mô hinh SSAN và mô hình của tác giả thì có độ đo AUC và độ chính xác (ACC) với ngưỡng 0.6 là gần tương đương nhau. Tuy nhiên, ta thấy FACL cải thiện so với SSAN ngay cả với ngưỡng tốt nhất của SSAN (0.7) với độ chính xác là 0.91. Trong tương lai, nhóm tác giả sẽ cố gắng cải thiện để mô hình của mình có kết quả vượt trội hơn hẳn các mô hình hiện đại hiện có.

3) Đánh giá về kích thước và tốc độ xử lý của mô hình: Ở phần này, tác giả thực hiện so sánh kích thước của ba mô hình lần lượt là ResNet-18, SSAN, và FACL (do chúng có kích cỡ chênh lệch nhau không quá lớn). Quan sát Bảng VI, chúng ta có thể thấy sự tương quan giữa số lượng tham số và tốc độ xử lý. Nếu số lượng tham số tăng lên, tốc độ xử lý cũng tăng lên, và ngược lại. Ví dụ, mô hình ResNet-18 có số lượng tham số thấp nhất là 11 triệu, đi kèm với tốc độ xử lý nhanh hơn cả là 0.025 giây. Ngược lại, mô hình SSAN có số lượng tham số lớn nhất với 31 triệu tham số với thời gian xử lý là 0.091 giây. Mô hình của tác giả cân bằng hơn hai mô trình trên cả về mặt hiệu năng cũng như có độ chính xác vượt trội.

V. KẾT LUẬN

Trong bài báo này, nhóm tác giả đã trình bày một phương pháp mới để phát hiện chống giả mạo khuôn mặt đã đạt được độ chính xác 91%. Nhóm tác giả đã thảo luận chi tiết về ba hướng khác nhau: dữ liệu, kiến trúc và khởi tạo, tổng hợp thành một giải pháp nhất quán, thể hiện những cải tiến đáng kể trên bộ thử nghiệm. Đầu tiên, tác giả đã chứng minh rằng việc lựa chọn cẩn thận một tập hợp con huấn luyện theo các loại mẫu giả mạo sẽ tổng quát hóa tốt hơn cho các cuộc tấn công không nhìn thấy được. Thứ hai, nhóm tác giả đã đề xuất một mô-đun tổng hợp sử dụng đầy đủ đặc trưng từ các phương thức khác nhau ở các cấp độ xử lý. Cuối cùng, chúng ta đã kiểm tra ảnh hưởng của việc truyền đặc trưng từ các mô hình được đào tạo trước khác nhau đối với tác vụ đích và cho thấy rằng việc sử dụng tập hợp các tác vụ liên quan đến khuôn mặt khác nhau khi thay nguồn domain để làm tăng tính ổn định và hiệu suất của hệ thống.

LỜI CẢM ƠN

Nghiên cứu này được tài trợ bởi đề tài cấp Bộ Thông tin & Truyền thông mã số: ĐT.25/23

REFERENCES

[1] G. Guo and N. Zhang, "A survey on deep learning based face recognition," Comput. Vis. Image Underst., vol. 189, p. 102805, 2019, doi: 10.1016/j.cviu.2019.102805.

[2] J. Cao, Y. Li, and Z. Zhang, "Celeb-500K: A large training dataset for face recognition," Proc. - Int. Conf. Image Process. ICIP, pp. 2406–2410, 2018, doi: 10.1109/ICIP.2018.8451704.

[3] A. Anwar and A. Raychowdhury, "Masked Face Recognition for Secure Authentication," pp. 1–8, 2020, [Online]. Available: http://arxiv.org/abs/2008.11104

Method	P&C&U to Z		P&C&Z to U		P&U&Z to C		C&U&Z to P	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
D2 AM	12.54	94.39	20.71	84.59	15.19	90.06	**15.09**	89.80
SDA	15.21	90.51	24.13	83.29	15.42	88.85	22.78	83.20
DRDG	12.28	94.59	18.79	87.63	15.34	90.54	15.42	90.50
ResNet-18	17.02	90.10	19.68	87.43	20.87	86.72	25.02	81.47
SSAN	10.68	**95.61**	17.62	88.21	15.80	89.93	15.46	**90.68**
Mô hình đề xuất - FACL	**10.28**	94.47	**16.26**	**90.68**	**13.80**	**94.27**	19.24	88.08

Table II
KẾT QUẢ KIỂM TRA GIỮA CÁC BỘ DỮ LIỆU

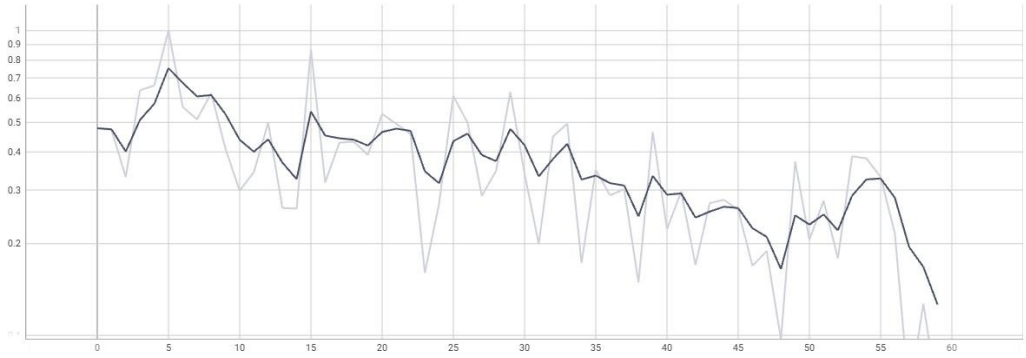

Figure 5. Biểu đồ hàm mất mát theo 60 epoch.

	Số lượng tham số (triệu)	Tốc độ xử lý (giây)
ResNet-18	11,7	0.025
SSAN	31,5	0.091
FACL	22,82	0.048

Table VI
KẾT QUẢ ĐO VỀ KÍCH THƯỚC CŨNG NHƯ TỐC ĐỘ XỬ LÝ CỦA MÔ HÌNH

	AUC	ACC (ngưỡng = 0.6)
ResNet-18	0.8645	0.6748
SSAN	0.9111	0.8855
Mô hình đề xuất	0.9188	0.9207

Table V
KẾT QUẢ ĐỘ CHÍNH XÁC VÀ CHỈ SỐ AUC TƯƠNG ỨNG VỚI MỖI MÔ HÌNH.

	Dự đoán	
Nhãn thực	Live	Fake
Live	421 (TP)	36 (FN)
Fake	8 (FP)	24 (TN)

Table III
MA TRẬN NHẦM LẪN CỦA SSAN VỚI NGƯỠNG 0.7

	Dự đoán	
Nhãn thực	Live	Fake
Live	427 (TP)	30 (FN)
Fake	5 (FP)	27 (TN)

Table IV
MA TRẬN NHẦM LẪN CỦA FACL VỚI NGƯỠNG 0.6

[4] Y. Zhong and W. Deng, "Towards Transferable Adversarial Attack against Deep Face Recognition," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1452–1466, 2021, doi: 10.1109/TIFS.2020.3036801.

[5] A. Liu et al., "Cross-ethnicity face anti-spoofing recognition challenge: A review," IET Biometrics, vol. 10, no. 1, pp. 24–43, 2021, doi: 10.1049/bme2.12002.

[6] Farmanbar, M., Toygar, Ö. Spoof detection on face and palmprint biometrics. SIViP 11, 1253–1260 (2017). https://doi.org/10.1007/s11760-017-1082-y

[7] H. Chen, Y. Chen, X. Tian and R. Jiang, "A Cascade Face Spoofing Detector Based on Face Anti-Spoofing R-CNN and Improved Retinex LBP," in IEEE Access, vol. 7, pp. 170116-170133, 2019, doi: 10.1109/ACCESS.2019.2955383.

[8] R. Ganjoo and A. Purohit, "Anti-Spoofing Door Lock Using Face Recognition and Blink Detection," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 1090-1096, doi: 10.1109/ICICT50816.2021.9358795.

[9] P. Anthony, B. Ay and G. Aydin, "A Review of Face Anti-spoofing Methods for Face Recognition Systems," 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, 2021, pp. 1-9, doi: 10.1109/INISTA52262.2021.9548404.

[10] D. Sharma and A. Selwal, "A face anti-spoofing approach based on generic sequential model using scale invariant features," 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Pitesti, Romania, 2021, pp. 1-6, doi: 10.1109/ECAI52376.2021.9515179.

[11] S. Zhang et al., "CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing," in IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 2, pp. 182-193, April 2020, doi: 10.1109/TBIOM.2020.2973001.

[12] S. Jia, X. Li, C. Hu, G. Guo and Z. Xu, "3D Face Anti-Spoofing With Factorized Bilinear Coding," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 10, pp. 4031-4045, Oct. 2021, doi: 10.1109/TCSVT.2020.3044986.

[13] M. A. Younus and T. M. Hasan, "Abbreviated View of Deepfake Videos Detection Techniques," 2020 6th International Engineering Conference “Sustainable Technology and Development" (IEC), Erbil, Iraq, 2020, pp. 115-120, doi: 10.1109/IEC49899.2020.9122916.

[14] S. R. Chavan, S. S. Sherekar and V. M. Thakre, "Factors Related To The Improvement of Face Anti-Spoofing Detection Techniques With CNN Classifier," 2021 International Conference on Computational Intelligence and Computing Applications (ICCICA), Nagpur, India, 2021, pp. 1-4, doi: 10.1109/ICCICA52458.2021.9697292.

[15] S. Mondal, "Implementation of Human Face and Spoofing Detection Using Deep Learning on Embedded Hardware," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225495.

[16] U. Muhammad and A. Hadid, "Face Anti-spoofing using Hybrid Residual Learning Framework," 2019 International Conference on Biometrics (ICB), Crete, Greece, 2019, pp. 1-7, doi: 10.1109/ICB45273.2019.8987283.

[17] R. Koshy and A. Mahmood, "Enhanced Anisotropic Diffusion-based CNN-LSTM Architecture for Video Face Liveness Detection," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2020, pp. 422-425, doi: 10.1109/ICMLA51294.2020.00074.

[18] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson and S. Z. Li, "Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 578-593, 2020, doi: 10.1109/TIFS.2019.2922241.

[19] Wang, Zhuo and Wang, Zezheng and Yu, Zitong and Deng, Weihong and Li, Jiahong and Li, Size and Wang, Zhongyuan, “Domain Generalization via Shuffled Style Assembly for Face Anti-Spoofing”, CVPR, 2022.

[20] Yiyou Sun and Yaojie Liu and Xiaoming Liu and Yixuan Li and Wen-Sheng Chu, “Rethinking Domain Generalization for Face Anti-spoofing: Separability and Alignment”, CVPR, 2023.

[21] W. Wang, P. Liu, H. Zheng, R. Ying and F. Wen, "Domain Generalization for Face Anti-Spoofing via Negative Data Augmentation." in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 2333-2344, 2023, doi: 10.1109/TIFS.2023.3266138.

[22] Zhekai Du, Jingjing Li, Lin Zuo, Lei Zhu, Ke Lu. "Energy-Based Domain Generalization for Face Anti-Spoofing". MM ’22: Proceedings of the 30th ACM International Conference on Multimedia. October 2022. Pages 1749–1757. https://doi.org/10.1145/3503161.3548073.

[23] Rossler, Andreas, et al. "Faceforensics++: Learning to detect manipulated facial images." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[24] Xun Huang and Serge Belongie, “Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization”, ICCV, 2017.

CONTRASTIVE LEARNING AND FEATURE SYNTHESIS APPROACH FOR GENERALIZED DOMAIN ANTI-SPOOFING IN FACE RECOGNITION

Abstract: "Face Anti-Spoofing (FAS) is an essential method in facial recognition systems to safeguard and accurately identify individuals. In recent years, facial spoofing detection algorithms have shown significant advancements, even in cases where model training data is limited. However, these deep learning algorithms are still relatively basic, and in many instances, they fail to detect facial spoofing. Recently, some models have leveraged pixel-level signals to address the FAS task. In this paper, we propose a novel evaluation method named FACL (Feature Aggregation and Contrastive Learning) based on the informative features of input facial images and contrastive learning. Our experiments conducted on two datasets: (1) the PTIT dataset and (2) the Zalo dataset, deliver superior and more effective results compared to several existing methods. Furthermore, our research also demonstrates the effectiveness of models learning different pixel patterns and simultaneously providing in-depth information for enhanced facial anti-spoofing supervision."

Keywords: face anti-spoofing, liveness detection, deep learning.

Bui Quoc Bao nhận bằng kỹ sư chuyên ngành Toán ứng dụng và Tin học tại Đại học Bách khoa Hà Nội năm 2021. Hiện tại, Kỹ sư Bảo đang là học viên Thạc sỹ tại Đại học Bách khoa Hà Nội. Từ năm 2021, Kỹ sư Bảo là chuyên viên phát triển và nghiên cứu Thị giác máy tính tại Trung tâm Công nghệ thông tin MobiFone. Hướng nghiên cứu của kỹ sư Bảo bao gồm nhận dạng ảnh, học sâu và học tăng cường.

Tran Anh Dat nhận bằng kỹ sư An toàn thông tin tại Học viện Công nghệ Bưu chính Viễn thông năm 2020 và bằng thạc sỹ Khoa học Máy tính tại Học viện Công nghệ Bưu chính Viễn thông, Việt Nam năm 2022. Hiện tại, Thạc sỹ Đạt đang là nghiên cứu sinh ngành Khoa học Máy tính tại Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Hướng nghiên cứu của thạc sỹ Đạt bao gồm nhận dạng ảnh và học sâu.

Vu Hoai Nam nhận bằng kỹ sư Điện tử Viễn thông tại Đại học Bách Khoa Hà Nội năm 2013 và bằng thạc sỹ Khoa học Máy tính tại Đại học Quốc gia Chonnam, Hàn Quốc năm 2015. Hiện tại, Thạc sỹ Nam đang là nghiên cứu sinh, đồng thời là giảng viên nghành Khoa học Máy tính tại Học viện Công nghệ Bưu chính Viễn thông. Hướng nghiên cưu của thạc sỹ Nam bao gồm xử lý ảnh UAV, học máy, và học sâu.

Nguyen Khanh Hung Hiện tại, Hưng đang là học Kỹ sư năm cuối tại Đại học Bách khoa Hà Nội. Và hiện đang là thực tập sinh phát triển và nghiên cứu Thị giác máy tính tại Trung tâm Công nghệ thông tin MobiFone. Hướng nghiên cứu của Hưng bao gồm nhận dạng ảnh, học sâu, học tăng cường và xử lý trên Edge deivce.

Vu Van Thuong nhận bằng kỹ sư và thạc sỹ Toán Tin tại Đại học Bách Khoa Hà Nội năm 2014 và năm 2022. Từ năm 2023, thạc sỹ Thương là giảng viên Viện Khoa học Kỹ thuật Bưu điện, Học viện Công nghệ Bưu chính Viễn thông. Hướng nghiên cứu của thạc sỹ Thương bao gồm xử lý ảnh, học máy, toán tối ưu và phân tích dữ liệu

Nguyen Viet Hung tốt nghiệp thạc sĩ năm 2009 tại ĐH Bách Khoa Grenoble và bảo vệ luận án Tiến sỹ năm 2013 tại đại học Rennes 1, Cộng Hòa Pháp. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: hệ thống thông tin vô tuyến thế hệ mới.