### CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ BÀI TOÁN

#### 1.1. Tổng quan về bài toán chống giả mạo khuôn mặt

Bài toán chống giả mạo khuôn mặt là một trong những ứng dụng quan trọng của công nghệ nhận dạng khuôn mặt. Kỹ thuật chống giả mạo khuôn mặt (anti-spoofing) được sử dụng để phát hiện và ngăn chặn các kỹ thuật giả mạo khuôn mặt, như sử dụng hình ảnh, video hoặc mô hình 3D của một người khác để qua mặt hệ thống nhận dạng khuôn mặt.

Trong bối cảnh hiện nay, khi công nghệ nhận dạng khuôn mặt được sử dụng rộng rãi trong các lĩnh vực như an ninh, giao thông hay quản lý điểm danh, việc sử dụng các kỹ thuật giả mạo khuôn mặt đã trở nên phổ biến. Điều này đã đặt ra một thách thức lớn cho các hệ thống nhận dạng khuôn mặt, khi cần đảm bảo tính chính xác và độ tin cậy của quá trình nhận dạng.

Để giải quyết vấn đề này, các kỹ thuật anti-spoofing được sử dụng để phát hiện các kỹ thuật giả mạo khuôn mặt bằng cách phân tích các đặc trưng khác nhau của khuôn mặt, chẳng hạn như độ sáng, độ tương phản, chuyển động và chiều sâu. Các phương pháp anti-spoofing cũng có thể sử dụng các thuật toán học máy để phát hiện các đặc trưng giả mạo khuôn mặt dựa trên các mẫu đào tạo.

Tuy nhiên, các kỹ thuật anti-spoofing vẫn đang phát triển và chưa hoàn thiện, đặc biệt là khi phải đối mặt với các kỹ thuật giả mạo khuôn mặt mới. Do đó, việc nghiên cứu và phát triển các kỹ thuật anti-spoofing còn rất cần thiết để đảm bảo tính chính xác và độ tin cậy của quá trình nhận dạng khuôn mặt.

Trong đề tài này, nhóm chúng em thực hiện nghiên cứu các kỹ thuật chống giả mạo và ứng dụng vào hệ thống điểm danh tự động dựa trên khuôn mặt. Dữ liệu đầu vào là các video stream thu được từ các camera giám sát tại các phòng. Các kỹ thuật được sử dụng bao gồm: face recognition (nhận dạng khuôn mặt), face detection (phát hiện khuôn mặt), face verification (xác minh khuôn mặt), face anti spoofing (chống giả mạo khuôn mặt) , dựa trên công nghệ học sâu (Deep Learning) và xử lý hình ảnh (Computer Vision). Kết quả nghiên cứu dự kiến bao gồm: Tự động điểm danh khuôn mặt trên video, lưu lại thông tin của thí sinh vào hệ thống của trường. Trong tương lai gần sau khi thực hiện nghiên cứu này nhóm sẽ xây dựng hệ thống phát hiện gian trong thời gian thực và có thể áp dụng trực tiếp tại các phòng thông qua camera giám sát.

#### 1.2. Một số kỹ thuật giải quyết bài toán

Có nhiều kỹ thuật anti-spoofing được sử dụng để giải quyết bài toán chống giả mạo khuôn mặt, tùy thuộc vào từng ứng dụng cụ thể. Dưới đây là một số kỹ thuật tiêu biểu:

- Phát hiện chuyển động: Kỹ thuật này sử dụng các cảm biến chuyển động để phát hiện các chuyển động giả mạo, chẳng hạn như việc sử dụng một video đã được ghi lại trước đó để qua mặt hệ thống nhận dạng khuôn mặt.

- Phát hiện độ sáng và độ tương phản: Kỹ thuật này sử dụng các cảm biến độ sáng và độ tương phản để phát hiện các hình ảnh giả mạo, chẳng hạn như việc sử dụng một bức ảnh của một người khác để qua mặt hệ thống nhận dạng khuôn mặt.

- Phân tích tín hiệu 3D: Kỹ thuật này sử dụng các cảm biến chiều sâu để phân tích các tín hiệu 3D của khuôn mặt và phát hiện các kỹ thuật giả mạo khuôn mặt, chẳng hạn như việc sử dụng mô hình 3D của một người khác để qua mặt hệ thống nhận dạng khuôn mặt.

- Phân tích động tác khuôn mặt: Kỹ thuật này sử dụng các cảm biến chuyển động để phân tích các động tác khuôn mặt và phát hiện các kỹ thuật giả mạo khuôn mặt, chẳng hạn như việc sử dụng một video được tạo ra bằng cách tạo ra các động tác khuôn mặt giả để qua mặt hệ thống nhận dạng khuôn mặt.

- Sử dụng các thuật toán học máy: Kỹ thuật này sử dụng các thuật toán học máy để phân tích các đặc trưng khuôn mặt và phát hiện các kỹ thuật giả mạo khuôn mặt dựa trên các mẫu đào tạo.

- Kết hợp các kỹ thuật: Các kỹ thuật anti-spoofing thường được kết hợp để cải thiện hiệu quả phát hiện giả mạo khuôn mặt. Ví dụ, sử dụng cả phân tích động tác khuôn mặt và phân tích tín hiệu 3D có thể cung cấp thông tin về hình dạng và độ sâu của khuôn mặt, giúp đảm bảo tính chính xác của hệ thống nhận dạng khuôn mặt. Kỹ thuật phân tích tín hiệu 3D có thể phát hiện giả mạo bằng cách phân tích các tín hiệu không gian 3D của khuôn mặt, trong khi kỹ thuật phân tích động tác khuôn mặt có thể phát hiện giả mạo bằng cách phân tích các chuyển động trên khuôn mặt.

- Các kỹ thuật anti-spoofing thường được kết hợp để cải thiện hiệu quả phát hiện giả mạo khuôn mặt. Ví dụ, sử dụng cả phân tích động tác khuôn mặt và phân tích tín hiệu 3D có thể cung cấp thông tin về hình dạng và độ sâu của khuôn mặt, giúp đảm bảo tính chính xác của hệ thống nhận dạng khuôn mặt. Kỹ thuật phân tích tín hiệu 3D có thể phát hiện giả mạo bằng cách phân tích các tín hiệu không gian 3D của khuôn mặt, trong khi kỹ thuật phân tích động tác khuôn mặt có thể phát hiện giả mạo bằng cách phân tích các chuyển động trên khuôn mặt.

- Sử dụng các thuật toán deep learning cũng là một kỹ thuật phổ biến để giải quyết bài toán chống giả mạo khuôn mặt. Các thuật toán này có thể học từ các mẫu đào tạo và phát hiện các kỹ thuật giả mạo khuôn mặt dựa trên các đặc trưng của khuôn mặt. Tuy nhiên, để đạt hiệu quả cao, cần có đủ dữ liệu đào tạo và các thuật toán học máy phải được huấn luyện đúng cách.

- Ngoài ra, việc sử dụng các hệ thống đa lớp để phân tích và xác định tính hợp lệ của dữ liệu đầu vào, cũng là một kỹ thuật phổ biến để giải quyết bài toán chống giả mạo khuôn mặt. Hệ thống đa lớp sẽ phân tích các thông tin từ nhiều nguồn khác nhau, như hình ảnh, âm thanh, và động tác, để xác định tính hợp lệ của khuôn mặt.

- Tóm lại, để giải quyết bài toán chống giả mạo khuôn mặt, cần sử dụng các kỹ thuật anti-spoofing khác nhau và kết hợp chúng để đạt hiệu quả cao nhất. Các kỹ thuật này có thể được sử dụng độc lập hoặc kết hợp với nhau để đảm bảo tính chính xác và độ tin cậy của hệ thống nhận dạng khuôn mặt.

#### 1.3. Phạm vi của đề tài

Phạm vi của bài toán chống giả mạo khuôn mặt rất rộng và đa dạng, tuỳ thuộc vào mục đích và điều kiện ứng dụng. Các phương pháp tiếp cận thường được chia thành các nhóm chính như: phân biệt giữa khuôn mặt thật và giả, phát hiện khuôn mặt giả, xác minh danh tính, nhận dạng khuôn mặt.

Trong phạm vi của đề tài này, chúng ta sẽ tập trung vào phân loại khuôn mặt thật và giả thông qua việc sử dụng kỹ thuật Anti-Spoofing. Anti-Spoofing là một kỹ thuật được sử dụng để phát hiện và ngăn chặn các mẹo giả mạo khuôn mặt thông qua việc xác định tính hợp lệ của dữ liệu khuôn mặt được cung cấp cho hệ thống.

Để đảm bảo tính hiệu quả và độ chính xác của hệ thống chống giả mạo khuôn mặt, chúng ta sẽ chỉ sử dụng các loại camera thông thường, bao gồm các loại camera giám sát, máy tính bảng, điện thoại, webcam máy tính. Chúng ta sẽ không sử dụng các thiết bị chuyên dụng như máy quét khuôn mặt hay máy quét vân tay để đảm bảo tính khả dụng và ứng dụng rộng rãi của hệ thống.

Mỗi lần chỉ nhận diện một khuôn mặt, mẫu được đặt trước camera không quá hai mét để đảm bảo độ chính xác của hệ thống. Việc giới hạn khoảng cách này sẽ giúp chúng ta đảm bảo rằng hệ thống có đủ dữ liệu để xác định tính hợp lệ của khuôn mặt mà không bị ảnh hưởng bởi những yếu tố như độ sáng, góc chụp hay khoảng cách quá xa.

#### 1.4. Các bài toán con cần giải quyết

Để đạt được mục tiêu chung của đề tài, chúng tôi sẽ tập trung vào giải quyết các bài toán con sau:

Nhận dạng khuôn mặt chính xác: Để đảm bảo tính chính xác của hệ thống điểm danh tự động dựa trên khuôn mặt, bước đầu tiên là cần phải nhận dạng khuôn mặt của người sử dụng với độ chính xác cao nhất có thể. Đây là bài toán cơ bản và quan trọng nhất cần giải quyết để đạt được mục tiêu chính của đề tài.

Phát hiện giả mạo khuôn mặt: Với các phương pháp giả mạo như in ảnh, dán ảnh, video, hoặc mặt nạ giả, hệ thống điểm danh tự động dựa trên khuôn mặt dễ dàng bị đánh lừa. Do đó, chúng tôi cần phải sử dụng các kỹ thuật Anti-Spoofing để phát hiện và ngăn chặn các phương pháp giả mạo này.

Xử lý ảnh và video: Để nhận dạng khuôn mặt và phát hiện giả mạo khuôn mặt, chúng tôi cần phải sử dụng các kỹ thuật xử lý ảnh và video như lọc, cắt, phóng to hoặc thu nhỏ hình ảnh, đồng bộ hóa video, giảm nhiễu, cân bằng ánh sáng, và nhiều kỹ thuật khác để đảm bảo tính chính xác và độ tin cậy của hệ thống.

Tối ưu hệ thống: Hệ thống điểm danh tự động dựa trên khuôn mặt có thể phức tạp và yêu cầu nhiều tài nguyên tính toán. Chúng tôi cần phải tối ưu hóa hệ thống để đảm bảo hiệu quả và tốc độ xử lý tối đa với số lượng tài nguyên tối thiểu.

Bảo mật dữ liệu: Dữ liệu khuôn mặt là nhạy cảm và đòi hỏi sự bảo mật cao. Chúng tôi cần phải thiết kế và triển khai các giải pháp bảo mật để đảm bảo dữ liệu của người sử dụng được bảo vệ và tránh những nguy cơ bị đánh cắp hoặc lộ ra ngoài.

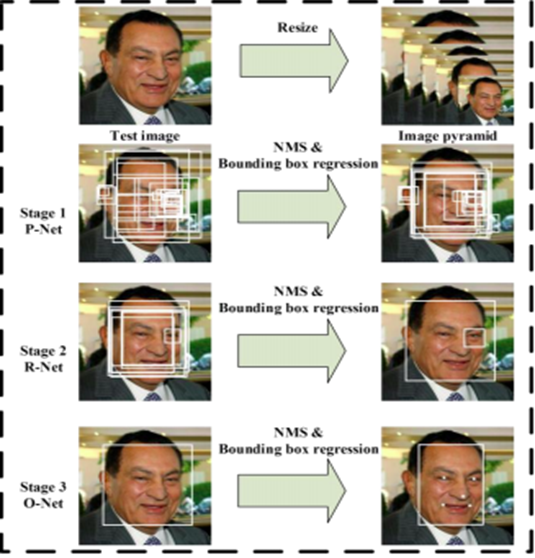
### CHƯƠNG 2: CÁC NGHIÊN CỨU CÓ LIÊN QUAN

#### 2.1. Multi-Task Cascaded Convolutional Networks (MTCNN)

MTCNN là một thuật toán sử dụng Deep Learning để phát hiện khuôn mặt trong ảnh, với độ chính xác cao và tốc độ xử lý nhanh. Có thể nói đây là phương pháp tiên phong trong việc kết hợp face detection và face alignment.

Mục đích của MTCNN là phát hiện khuôn mặt trong ảnh và trích xuất các điểm đặc trưng của khuôn mặt, bao gồm các điểm góc mắt, mũi và miệng. Điểm mạnh của MTCNN là khả năng phát hiện được khuôn mặt ở nhiều kích cỡ khác nhau và trong các điều kiện ánh sáng khác nhau.

MTCNN sử dụng một mô hình CNN (Convolutional Neural Network) với nhiều tầng tích chập và tầng max-pooling để trích xuất các đặc trưng của khuôn mặt. Mô hình này được chia thành 3 mạng con (P-Net, R-Net và O-Net) để phát hiện khuôn mặt ở các kích cỡ khác nhau và trích xuất các điểm đặc trưng của khuôn mặt.

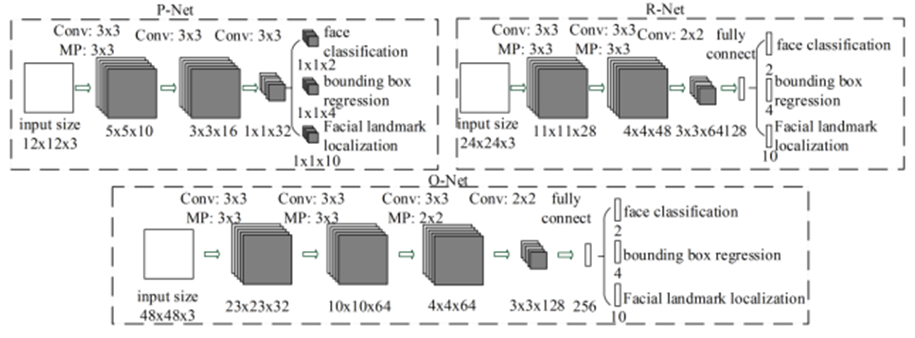
[](#D2L_fig_label_Tổng quan MTCNN)  
Figure 1: Tổng quan MTCNN

Mạng Multi-Task Cascaded Convolutional Networks (MTCNN) bao gồm 3 thành phần chính, bao gồm:

1. P-Net (Proposal Network): Thành phần đầu tiên của MTCNN, nó là một mạng tích chập (CNN) nhằm phát hiện và tạo ra các khuôn mặt ứng viên (candidates) dựa trên các tính năng của ảnh đầu vào. P-Net hoạt động với nhiều kích thước khác nhau để phát hiện được các khuôn mặt nhỏ và lớn trong ảnh.

2. R-Net (Refine Network): Thành phần thứ hai của MTCNN, nó được sử dụng để loại bỏ các ứng viên không phải là khuôn mặt và chuyển các ứng viên còn lại thành các ứng viên tốt hơn. R-Net cũng là một mạng tích chập (CNN), có cấu trúc phức tạp hơn so với P-Net và sử dụng đầu vào là các ảnh đã được cắt chính xác từ P-Net.

3. O-Net (Output Network): Thành phần cuối cùng của MTCNN, nó được sử dụng để xác định chính xác các đặc trưng của khuôn mặt, bao gồm vị trí, kích thước và hướng của khuôn mặt. O-Net cũng là một mạng tích chập (CNN) và sử dụng các đặc trưng được trích xuất từ R-Net. Nó cũng có cấu trúc phức tạp hơn R-Net và sử dụng các tầng fully connected layer để tính toán các thông số của khuôn mặt.

[](#D2L_fig_label_Chi tiết kiến trúc các mạng trong mô hình)  
Figure 2: Chi tiết kiến trúc các mạng trong mô hình

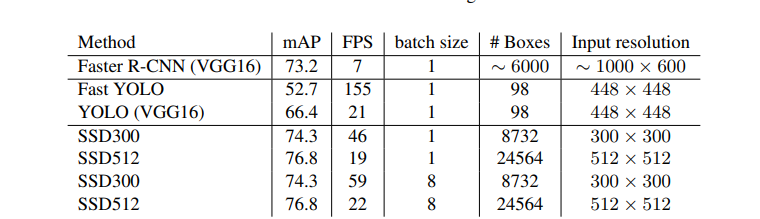
Các thành phần chính của MTCNN được thiết kế để làm việc liên tục với nhau, bằng cách chuyển các khuôn mặt ứng viên từ P-Net đến R-Net và sau đó đến O-Net. Các thành phần này được cấu hình sao cho có thể xử lý ảnh đầu vào với nhiều kích thước khác nhau và phát hiện được các khuôn mặt trong nhiều điều kiện ánh sáng và phông nền khác nhau. MTCNN là một trong những mô hình phát hiện khuôn mặt đáng tin cậy nhất hiện nay và được sử dụng rộng rãi trong các ứng dụng thực tế.

#### 2.2. Single Shot MultiBox Detector (SSD)

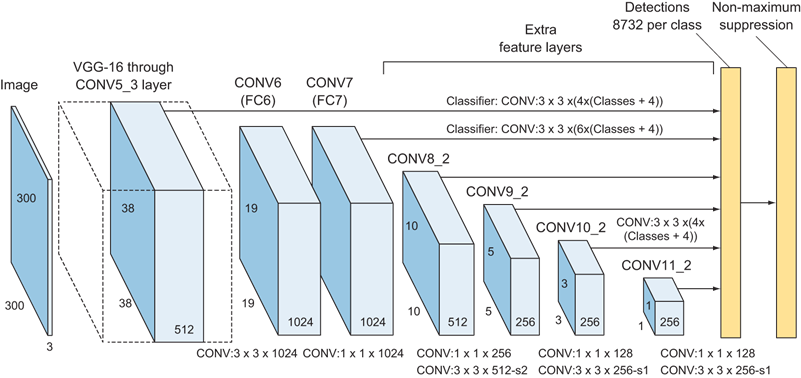
##### 2.2.1 Tổng quan về SSD

SSD (Single Shot Multibox Detector) là một mô hình phát hiện đối tượng dựa trên Deep Learning được giới thiệu bởi Wei Liu, Dragomir Anguelov, và cộng sự của họ vào năm 2016. Mô hình này được phát triển để giải quyết vấn đề phát hiện đối tượng trên các ảnh độ phân giải cao với tốc độ xử lý nhanh hơn so với các mô hình trước đó.

SSD đã được đánh giá hiệu quả trong việc phát hiện đối tượng trên các bộ dữ liệu benchmark, đặc biệt là trong việc phát hiện đối tượng đa dạng về kích thước và độ phân giải. Với kết quả đạt được, SSD được xem là một trong những mô hình phát hiện đối tượng tiên tiến nhất hiện nay.

[](#D2L_fig_label_Bảng so sánh tốc độ xử lý và độ chính xác của các lớp model object detection)  
Figure 3: Bảng so sánh tốc độ xử lý và độ chính xác của các lớp model object detection

##### 2.2.2 Kiến trúc của ssd



Kiến trúc của SSD (Single Shot MultiBox Detector) bao gồm các thành phần chính sau đây:

* Mạng trích xuất đặc trưng (Feature Extraction Network): Một mạng tích chập được sử dụng để trích xuất các đặc trưng của ảnh đầu vào. Kiến trúc của mạng này thường được lựa chọn từ các kiến trúc mạng tích chập đã được huấn luyện trước đó như VGG, ResNet, MobileNet, v.v. Như trên hình chúng ta có thể thấy nhóm tác giả sử dụng mạng VGG16 đã được bỏ các lớp fully connected layers.
* Multi-scale feature layers: series convolution filters được thêm vào sau base network, bắt đầu từ CONV6 như hình bên dưới. Những layers này sẽ giảm kích thước để cho phép dự đoán objects với nhiều scales khác nhau và giảm kích thước đầu vào ở các Conv layer tiếp theo.
* Non-maximum suppression - NMS được dùng để loại bỏ overlapping boxes.

Với các thành phần này, SSD có thể phát hiện và phân loại các vật thể trong ảnh chỉ trong một lần feed-forward (single shot) và được sử dụng rộng rãi trong các ứng dụng thị giác máy tính và trí tuệ nhân tạo.

##### 2.2.3 Loss Function

**Matching strategy**: Trong suốt quá trình training chúng ta cần xác định default boxes khớp với ground-truth box. Đối với mỗi ground-truth box chúng ta chọn các default boxes có IoU lớn hơn 0.5 (coi là positive examples). Điều này làm đơn giản vấn đề, nó cho phép model có thể dự đoán được nhiều boxes có score cao mà overlap với nhau thay vì chỉ lấy duy nhất một box có overlap lớn nhất.

#### 2.3. RetinaNet

RetinaNet là một kiến trúc mạng neural được đề xuất bởi Facebook AI Research vào năm 2017 để giải quyết bài toán phát hiện đối tượng. Kiến trúc này sử dụng phương pháp đặc trưng trên các cấp độ khác nhau để cải thiện độ chính xác của mô hình.

RetinaNet sử dụng một mạng neural tích chập (CNN) để trích xuất các đặc trưng từ ảnh đầu vào. Sau đó, mô hình sử dụng một phần tử con gọi là "feature pyramid network" để lấy các đặc trưng từ các cấp độ khác nhau trong mạng để phát hiện các đối tượng có kích thước khác nhau.

RetinaNet sử dụng một phương pháp mới gọi là "Focal Loss" để giải quyết vấn đề mất cân bằng giữa các lớp đối tượng trong quá trình huấn luyện. Phương pháp này giúp tập trung vào các trường hợp khó phân loại hơn, tăng độ chính xác của mô hình.

RetinaNet đã đạt được kết quả tốt trong các cuộc thi phát hiện đối tượng trên các tập dữ liệu lớn như COCO và PASCAL VOC. Kiến trúc này cũng đã được áp dụng trong nhiều ứng dụng thực tế như phát hiện bàn tay trong xe hơi tự lái, phát hiện đối tượng trên không gian đô thị, phát hiện người dùng trong ảnh chụp selfie, và nhiều hơn nữa.

#### 2.4. Bài toán Face Recognition

Bài toán Face Recognition là một trong những bài toán quan trọng trong lĩnh vực xử lý ảnh và nhận dạng. Nói chung, bài toán này là một quá trình xác định xem một bức ảnh có chứa khuôn mặt của một người nào đó hay không, và nếu có, thì đó là khuôn mặt của ai.

Để giải quyết bài toán Face Recognition, các nhà nghiên cứu đã đưa ra nhiều phương pháp và thuật toán khác nhau. Trong đó, có những phương pháp dựa trên cấu trúc mô hình mạng neural như Convolutional Neural Networks (CNN), một số phương pháp khác dựa trên học sâu (deep learning) để xây dựng các mô hình đối tượng (object models) của khuôn mặt.

Một trong những phương pháp phổ biến nhất để giải quyết bài toán Face Recognition là sử dụng mô hình mạng neural CNN. Các mô hình này có khả năng học và rút trích các đặc trưng đại diện cho khuôn mặt, từ đó xác định và so sánh với các đặc trưng của các khuôn mặt khác. Các mô hình CNN thường được huấn luyện trên các bộ dữ liệu lớn để đảm bảo tính tổng quát và độ chính xác cao.

Một số mô hình CNN phổ biến trong bài toán Face Recognition là FaceNet, VGGFace, DeepID, và DeepFace. Trong đó, mô hình FaceNet là một trong những mô hình tiên tiến nhất, được đề xuất bởi Google vào năm 2015. FaceNet sử dụng một kiến trúc mạng neural sâu để rút trích các đặc trưng của khuôn mặt, từ đó tạo ra các vectơ đặc trưng để thực hiện việc nhận dạng.

Ngoài ra, các phương pháp khác cũng được đề xuất để giải quyết bài toán Face Recognition như Eigenface, Fisherface, Local Binary Pattern Histogram (LBPH), và Histogram of Oriented Gradients (HOG). Tuy nhiên, các phương pháp này thường có độ chính xác thấp hơn so với các mô hình CNN.

Tóm lại, bài toán Face Recognition là một bài toán rất quan trọng và đang được nghiên cứu rộng rãi trong lĩnh vực xử lý ảnh và nhận dạng. Các mô hình mạng neural CNN đang là một trong những phương pháp tiên tiến nhất hiện nay.

#### 2.5. Bài toán Face Anti Spoofing

Bài toán Face Anti-Spoofing là một trong những bài toán quan trọng trong lĩnh vực nhận diện khuôn mặt. Nó có vai trò quan trọng trong việc đảm bảo tính an toàn và bảo mật của các hệ thống sử dụng nhận diện khuôn mặt, đặc biệt là trong các ứng dụng tài chính, an ninh, quân sự, v.v.

Bài toán Face Anti-Spoofing nhằm phát hiện và ngăn chặn việc giả mạo khuôn mặt bằng các kỹ thuật chống lại các loại tấn công giả mạo khác nhau, bao gồm các kỹ thuật giả mạo vật thể, kỹ thuật giả mạo ảnh và video.

Các kỹ thuật Face Anti-Spoofing có thể được chia thành hai loại chính: kỹ thuật dựa trên đặc trưng và kỹ thuật dựa trên hành vi.

Kỹ thuật dựa trên đặc trưng tập trung vào việc phát hiện các đặc trưng khác biệt giữa khuôn mặt thật và khuôn mặt giả mạo, ví dụ như tập trung vào phát hiện các vật liệu giả tạo như silicon, giả da, v.v. hoặc phát hiện các đặc trưng khác như màn hình LCD, các phông nền, v.v. Các kỹ thuật này có thể sử dụng các thuật toán học máy như SVM, Random Forest, Neural Network, v.v.

Kỹ thuật dựa trên hành vi tập trung vào việc phát hiện các hành vi khác biệt giữa khuôn mặt thật và khuôn mặt giả mạo, ví dụ như phát hiện các hành vi giả mạo như di chuyển quá nhanh, không tự nhiên, v.v. Các kỹ thuật này có thể sử dụng các thuật toán học sâu như LSTM, CNN, v.v.

Các nghiên cứu về Face Anti-Spoofing đã đạt được nhiều tiến bộ trong những năm gần đây và các kỹ thuật mới đang được phát triển để đáp ứng yêu cầu ngày càng cao của bảo mật khuôn mặt.

## CHƯƠNG 3: PROPOSED ALGORITHMS

### 3.1 Face Detection & Facial Landmark Detection

#### 3.1.1 Giới thiệu thuật toán RetinaFace

RetinaFace là một thuật toán Face Detection được phát triển dựa trên RetinaNet,nó có khả năng xác định các khuôn mặt với độ chính xác cao và có thể nhận dạng khuôn mặt trong các điều kiện ánh sáng khác nhau, với độ phân giải cao và thậm chí là khi khuôn mặt bị che khuất bởi phụ kiện như kính đeo mắt hay khẩu trang. Đặc biệt, RetinaFace có khả năng nhận dạng khuôn mặt của nhiều đối tượng trong cùng một hình ảnh.

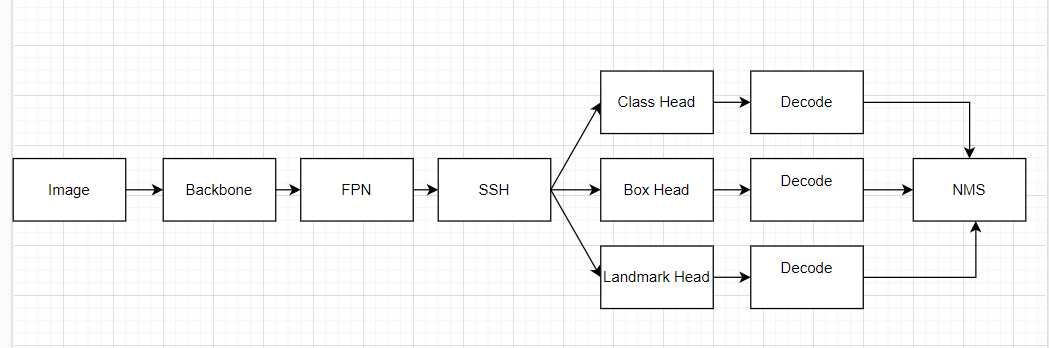
Thuật toán RetinaFace sử dụng một mạng neural network đa tầng để xác định các đặc trưng của khuôn mặt. Mạng neural network này có nhiều lớp convolutional layer để trích xuất các đặc trưng, sau đó kết hợp các đặc trưng này để tạo ra một bản đồ đặc trưng của khuôn mặt. Bản đồ đặc trưng này sau đó được sử dụng để xác định vị trí và kích thước của khuôn mặt trong hình ảnh.

RetinaFace cũng sử dụng một phương pháp mới để đo lường độ chính xác của các đối tượng nhận dạng, gọi là "focal loss". Phương pháp này giúp cải thiện khả năng phân loại của mô hình bằng cách tập trung vào các đối tượng khó nhận dạng.

Với khả năng nhận dạng khuôn mặt chính xác và đa dạng, RetinaFace được sử dụng rộng rãi trong các ứng dụng nhận dạng khuôn mặt, như nhận dạng khuôn mặt trong hình ảnh và video, xác thực khuôn mặt cho các ứng dụng an ninh và hệ thống giám sát, và trong các ứng dụng truyền thông xã hội như tạo ảnh đại diện và nhận dạng khuôn mặt trong ảnh chụp selfie.

#### 3.1.2 Cấu trúc thuật toán RetinaFace

Cấu trúc của RetinaFace gồm các phần chính: Backbone Network, Feature Pyramid Network,Single Stage Headless, subnet và Non-Maximum Suppression .



*Figure 6.*

*Figure 6.*

*Sơ đồ cấu trúc thuật toán RetinaFace*

* Backbone network: Backbone network của RetinaFace là một mạng CNN được sử dụng để trích xuất đặc trưng từ ảnh đầu vào. Một số kiến trúc mạng được sử dụng làm backbone network bao gồm ResNet50, ResNet101, ResNet152 và MobileNetV1.
* Feature Pyramid Network (FPN): FPN là một phần quan trọng trong RetinaFace, được sử dụng để tăng cường độ chính xác của việc dò tìm khuôn mặt. FPN được sử dụng để tạo ra một bộ đặc trưng (feature map) có nhiều tỷ lệ khác nhau cho mỗi vị trí trên ảnh đầu vào. Bộ đặc trưng này sẽ được sử dụng để dò tìm các đặc trưng khuôn mặt ở nhiều tỷ lệ khác nhau.
* Single Stage Headless(SSH):là một phần của mô hình CNN được sử dụng để trích xuất các đặc trưng từ hình ảnh đầu vào. SSH là một mô hình tích chập sâu, được thiết kế để tạo ra các đặc trưng chính xác từ hình ảnh.SSH được sử dụng để trích xuất các đặc trưng từ các vùng quan trọng của hình ảnh, như các khuôn mặt, mắt, mũi và miệng. Các đặc trưng này sau đó được sử dụng để phát hiện các khuôn mặt và các đối tượng khác trên hình ảnh
* Subnet: Subnet là một mạng CNN được sử dụng để xác định tọa độ và đặc trưng của khuôn mặt trên ảnh đầu vào. Trong RetinaFace, subnet được chia thành 3 phần: detection subnet, bbox regression subnet và landmark regression subnet. Mỗi phần sẽ được sử dụng để dò tìm các khuôn mặt trên ảnh đầu vào và xác định các thông tin về tọa độ, kích thước, hướng và các đặc trưng của khuôn mặt đó.
* Non-Maximum Suppression(NMS) trong RetinaFace là một kỹ thuật được sử dụng để giảm số lượng dự đoán trùng lặp trong quá trình phát hiện khuôn mặt. Khi áp dụng RetinaFace, mỗi khuôn mặt sẽ được dự đoán bởi nhiều khung hình (bounding box) với độ tin cậy khác nhau. Tuy nhiên, để giảm bớt số lượng khung hình dư thừa, NMS được sử dụng để loại bỏ các khung hình có độ tin cậy thấp hơn và giữ lại các khung hình có độ tin cậy cao hơn.Cụ thể, quá trình NMS trong RetinaFace được thực hiện bằng cách sắp xếp các khung hình theo độ tin cậy giảm dần và loại bỏ các khung hình có độ tương đồng (IoU) lớn hơn ngưỡng xác định. NMS là một bước quan trọng để giảm bớt số lượng khung hình trùng lặp và giúp tăng độ chính xác của quá trình phát hiện khuôn mặt.

#### 3.1.3 Cách hoạt động của thuật toán RetinaFace

Các bước thực hiện của thuật toán RetinaFace gồm:

* Chuẩn bị dữ liệu: Ảnh đầu vào cần được chuẩn bị trước khi đưa vào mô hình. Các bước chuẩn bị bao gồm: đọc ảnh, chuyển đổi ảnh sang dạng tensor và chuẩn hóa kích thước ảnh.
* Trích xuất đặc trưng: Mô hình RetinaFace sử dụng một backbone để trích xuất đặc trưng từ ảnh đầu vào. Backbone này thường được xây dựng dựa trên mạng CNN (Convolutional Neural Network) như ResNet hoặc MobileNet. Sau khi trích xuất đặc trưng, đầu ra sẽ là các feature map.
* Tạo đặc trưng với độ phân giải khác nhau:FPN giúp tạo ra một số lượng các feature map có độ phân giải khác nhau từ các tầng sâu của mạng CNN.Do kích thước đối tượng trong các feature map được tạo ra khác nhau, việc sử dụng chúng để phát hiện đối tượng trên toàn bộ ảnh sẽ không hiệu quả . Do đó FPN sử dụng một mạng Upsample và một mạng Shortcut Connection để tạo ra các feature map có độ phân giải cao hơn từ các feature map có độ phân giải thấp hơn. Quá trình này được thực hiện đệ quy cho đến khi đạt được các feature map có độ phân giải cao nhất.Các feature map có độ phân giải khác nhau này được sử dụng để phát hiện đối tượng với kích thước khác nhau trong ảnh đầu vào.Các feature map có độ phân giải cao hơn sẽ được sử dụng để phát hiện đối tượng nhỏ hơn trong khi các feature map có độ phân giải thấp hơn sẽ được sử dụng để phát hiện đối tượng lớn hơn
* Tính toán anchor boxes: Để tìm ra các vị trí ứng viên cho khuôn mặt trên ảnh, RetinaFace sử dụng anchor boxes. Anchor box là một hình chữ nhật với kích thước và tỷ lệ khác nhau, được đặt trên toàn bộ ảnh.
* Tính toán confidence score: Subnet tiếp theo sẽ tính toán confidence score cho từng anchor box, biểu thị mức độ có khuôn mặt trong anchor box đó.
* Tính toán offset: Sau đó, subnet tiếp theo sẽ tính toán offset để dịch chuyển anchor box đó để nó có chính xác vị trí của khuôn mặt.
* Áp dụng NMS: Các anchor box có confidence score cao sẽ được giữ lại. Các anchor box còn lại sẽ bị loại bỏ bằng phương pháp Non-Maximum Suppression (NMS).
* Trả về các khuôn mặt: Các anchor box được giữ lại sẽ được sử dụng để xác định vị trí và kích thước của khuôn mặt trên ảnh. RetinaFace sẽ trả về các bounding box chứa khuôn mặt và điểm tự tin (confidence score) tương ứng với mỗi khuôn mặt.



*Figure 7.*

*Figure 7.*

*Minh họa đầu ra thuật toán RetinaFace*

### 3.2 Face recognition

#### 3.2.1 Giới thiệu thuật toán ArcFace

ArcFace là một thuật toán được sử dụng trong face recognition để cải thiện độ chính xác và độ tin cậy của quá trình nhận dạng khuôn mặt. ArcFace là một mô hình học sâu (deep learning model) được đề xuất bởi cộng đồng nghiên cứu về trí tuệ nhân tạo vào năm 2018.

ArcFace cải tiến trên mô hình FaceNet (một mô hình face recognition phổ biến) bằng cách sử dụng hàm loss mới được gọi là ArcFace loss. Hàm loss này tạo ra một khoảng cách lớn hơn giữa các lớp khác nhau trong quá trình huấn luyện mô hình, giúp tăng độ phân biệt giữa các khuôn mặt và cải thiện độ chính xác của quá trình nhận dạng.

Một điểm mạnh của ArcFace là nó có khả năng nhận dạng các khuôn mặt ở nhiều góc độ và ánh sáng khác nhau, giúp nó trở thành một trong những phương pháp face recognition hiệu quả nhất trong thực tế.

ArcFace được sử dụng rộng rãi trong các ứng dụng nhận dạng khuôn mặt, bao gồm các hệ thống giám sát an ninh, ứng dụng di động, công cụ chứng thực người dùng và các ứng dụng trong lĩnh vực giải trí.

#### 3.2.2 Cấu trúc thuật toán ArcFace

Cấu trúc của thuật toán ArcFace gồm ba phần chính:

* Backbone network: Đây là một mạng nơ-ron sâu (deep neural network) được sử dụng để học các đặc trưng của khuôn mặt. Mạng nơ-ron này thường là một mạng tích chập (convolutional neural network - CNN) với nhiều lớp, được huấn luyện trên một tập dữ liệu lớn chứa các khuôn mặt. Một số backbone network thường được sử dụng cho ArcFace bao gồm ResNet, Inception, và MobileNet.
* ArcFace loss: Đây là một hàm mất mát được sử dụng để học cách tính toán khoảng cách giữa các khuôn mặt. Hàm mất mát này được thiết kế để đảm bảo rằng khoảng cách giữa các khuôn mặt cùng một người sẽ nhỏ hơn khoảng cách giữa các khuôn mặt của các người khác nhau. Điều này giúp cho việc phân loại và nhận diện khuôn mặt trở nên chính xác hơn.
* Softmax classifier: Sau khi các đặc trưng của khuôn mặt đã được học bởi backbone network và khoảng cách giữa các khuôn mặt đã được tính toán bởi ArcFace loss, thuật toán sử dụng một phân loại softmax để phân loại khuôn mặt vào các nhóm tương ứng với các cá nhân. Các nhóm này được xác định trước bằng cách đưa cho thuật toán một tập dữ liệu huấn luyện chứa các khuôn mặt đã được gán nhãn với tên của từng cá nhân.

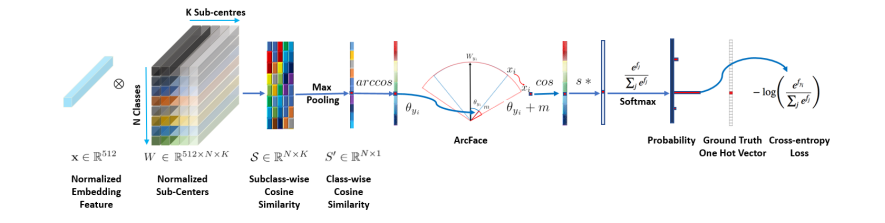
Tổng quan về cấu trúc của thuật toán ArcFace là:

* Sử dụng một backbone network để học các đặc trưng của khuôn mặt.
* Sử dụng Additive Angular Margin loss để học cách tính toán khoảng cách giữa các khuôn mặt.
* Sử dụng softmax classifier để phân loại khuôn mặt vào các nhóm tương ứng với các cá nhân.

#### 3.2.3 Cách hoạt động của thuật toán ArcFace

Cách hoạt động của ArcFace bao gồm các bước sau:

* Lấy đầu vào là một bức ảnh khuôn mặt, được mã hóa bằng một vector đặc trưng với chiều dài cố định.
* Tính toán vector đặc trưng này trên một mạng nơ-ron được huấn luyện trước.
* Vector đặc trưng này được chuẩn hoá để có độ dài bằng 1.
* Sử dụng ma trận trọng số W, chuyển đổi vector đặc trưng đó thành một vector mới trong không gian lớp (class space).
* Thực hiện chuẩn hóa lần nữa trên vector mới này.
* Tính toán cosine similarity giữa vector đặc trưng đã được chuẩn hoá và vector mới trong không gian lớp.
* Sử dụng hàm softmax trên cosine similarity để chuyển đổi nó thành một xác suất phân loại.
* Để đạt được sự tách biệt giữa các lớp, hàm mất mát ArcFace loss được tính toán bằng cách lấy sự khác biệt giữa cosine similarity của lớp đích và các lớp khác nhau, sau đó đưa ra một giá trị mất mát. Hàm mất mát này được tối ưu hóa thông qua thuật toán gradient descent để cải thiện độ chính xác của phân loại khuôn mặt.



*Figure 8.*

*Figure 8.*

*Sơ đồ minh họa thuật toán ArcFace*

3.3 Face Anti Spoofing

3.3.1 Một số phương pháp tấn công giả mạo

Có khá nhiều phương pháp để lừa một cái máy tính. Có thể phân ra thành, tấn công 2D và tấn công 3D:

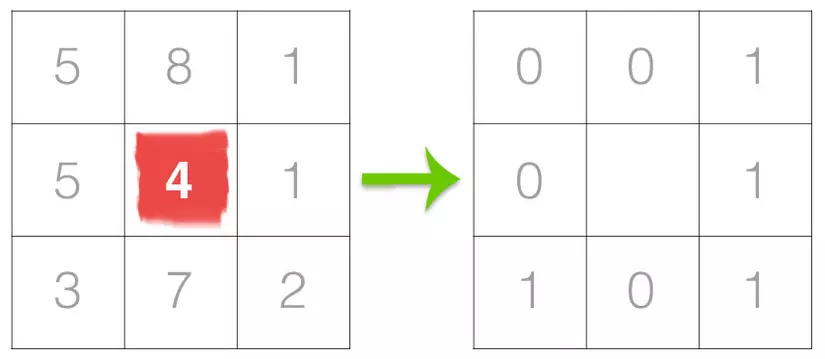
* 2D attacks: một tấm ảnh, một tấm hình in, một video chứa mặt người đó, những ảnh được thay liên tiếp để tạo sự chuyển động giả…
* 3D attacks: đeo mặt nạ giả, hình in 3D, tượng mặt người, makeup hoặc thậm chí là một con robot có khuôn mặt được điều chỉnh mô phỏng khuôn mặt ai đó.

3.3.2 Các phương pháp Face Anti-spoofing

3.3.2.1 Local binary pattern

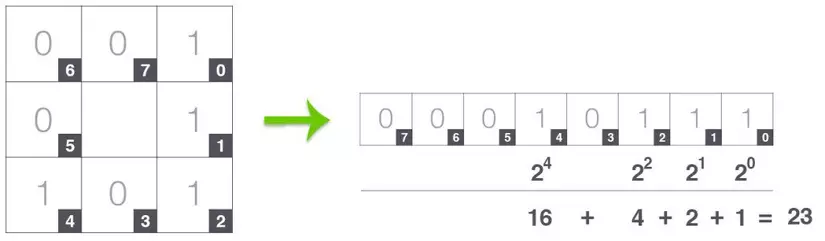
**Local Binary Pattern** (LBPs) tính toán biểu diễn cục bộ (Local representation). Local representation này được xây dựng bằng cách so sánh từng pixel với các pixel lân cận xung quanh của nó.

Local binary patten hoạt động như sau: Thuật toán chạy lần lượt chạy qua các điểm ảnh, với mỗi điểm ảnh thuật toán sẽ so sánh với các điểm ngay cạnh nó nếu lớn hơn nó các điểm xung quanh sẽ nhận giá trị là 1, nếu nhỏ hơn sẽ nhận giá trị là 0. Tiếp theo thuật toán dàn phẳng các điểm xung quanh được một số nhị phân, tiếp theo chuyển sang hệ 10 và cuối cùng lấy giá trị đó thay thế cho điểm ảnh hiện tại.



*Figure 9.*

*Figure 9.*



*Figure 10.*

*Figure 10.*

Ưu điểm: Phương pháp dễ dàng thực hiện, yêu cầu chi phí triển khai về phần cứng thấp.

Khuyết điểm: Độ chính xác mà phương pháp này mang lại phụ thuộc rất lớn vào điều kiện bên ngoài như ánh sáng, chất lượng camera, nên thường rất dễ bị vượt qua.

3.3.2.2 Eye Blink Detection

Trung bình con người nháy mắt 15 đến 30 lần mỗi phút. Thời gian chớp mắt của con người rơi vào khoảng 250 mm giây, trong khi đó một chiếc camera trung bình có thể bắt tới 30 khung hình trên giây tức rơi vào khoảng 50 mm. Điều này có nghĩa chúng ta hoàn toàn có thể dùng phương phát phát hiện nháy mắt để phát hiện một khuôn mặt có phải là thật hay từ một ảnh in ra. Tuy nhiên phương pháp này có nhiều nhược điểm điển hình như kẻ giả mạo quay một video sau đó phát lại để đánh lừa hệ thống.

3.3.2.3 CNN

Việc CNN ra đời đã tạo nên một cuộc cánh mạng lớn đối với ngành AI nói chung và Computer Vision nói riêng. Từ khi CNN ra đời nó đã giải quyết rất nhiều loại bài toán khác nhau với độ chính xác cao. Để ý chúng ta cũng có thể coi việc xác định khuôn mặt thật giả cũng giống như nhưng bài toán phân loại nhị phân khác: phân biệt chó mèo, phân biệt nam nữ... Chúng có thể xây dựng một bộ dữ liệu gồm hai nhãn khuôn mặt thật giả sau đó đưa vào các mạng CNN nổi tiếng để tiến hành huấn luyện phân loại.

Đây là một hướng đi nhiều tiềm năng và trên thế gới có nhiều paper đi theo hướng này, tuy nhiên bài toán phát hiện khuôn mặt thật giả là một bài toán khó, kẻ giả mạo có nhiều cách tấn công nên trên thế gới cũng chia ra làm nhiều hướng giải quyết với nhưng ưu nhược điểm khác nhau. Trong quá trình tìm hiểu nhóm xin liệt kê một số phương pháp nổi tiếng dựa trên mạng CNN dưới đây.