Nhận Diện Khuôn Mặt

Giảng viên hướng dẫn: TS. Mai Tiến Dũng
Sinh viên thực hiện:
Nguyễn Công Nguyên - 21521200
Lê Tiến Quyết - 21520428

CS331.P11
Khoa Khoa học máy tính
Trường Đại học Công nghệ thông tin - ĐHQG HCM

Ngày 20 tháng 2 năm 2025

Nội dung

Lý do chọn bài toán

- **Tầm quan trọng:** Nhận diện khuôn mặt là lĩnh vực quan trọng trong trí tuệ nhân tạo, ứng dụng rộng rãi trong an ninh và thương mại.
- **Ứng dụng:** Kiểm soát truy cập và phân tích hành vi người dùng.
- **Thách thức:** Đa dạng góc nhìn, ánh sáng và xử lý giả mạo.
- **Mục tiêu:** Xây dựng hệ thống nhận diện khuôn mặt chính xác và hiệu quả.

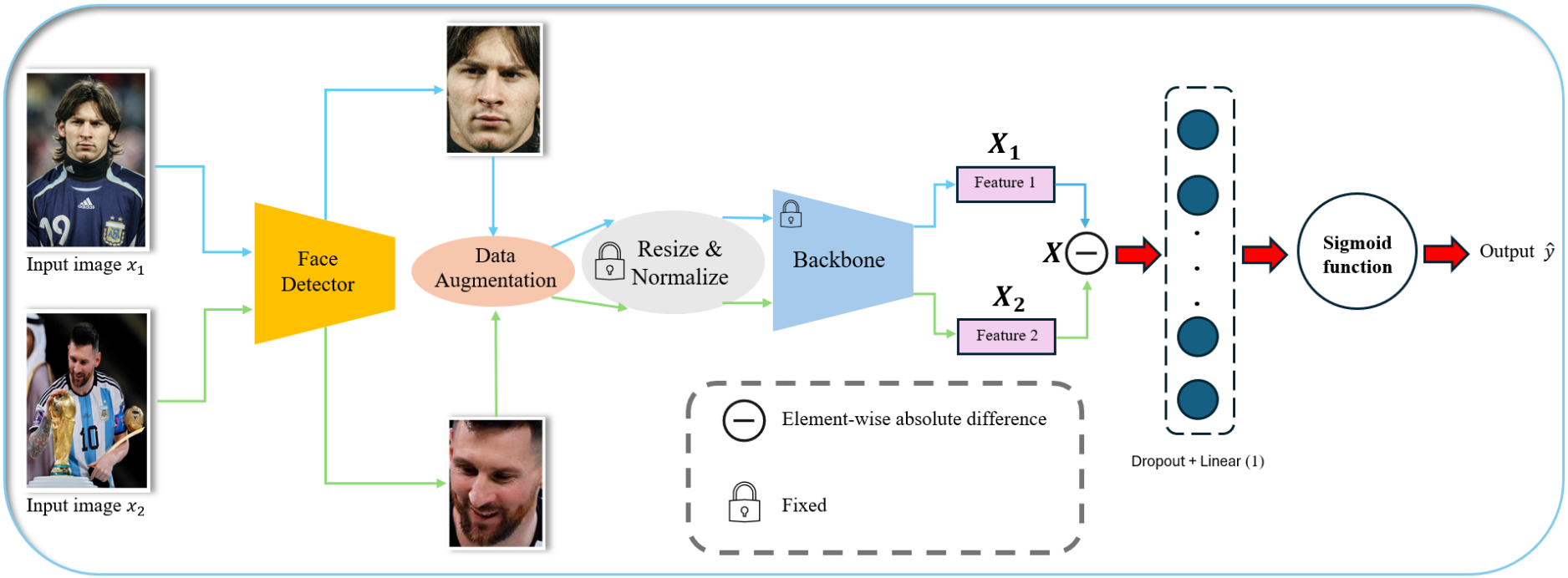
Hình 1: Ứng dụng của nhận diện khuôn mặt

Phát biểu bài toán

- **Input:**
  - Một bộ dữ liệu D gồm N ảnh, mỗi ảnh chỉ chứa một khuôn mặt duy nhất đã được xác định danh tính: $D = \{(I_1, y_1), (I_2, y_2), \ldots, (I_N, y_N)\}$. Với y_i là danh tính của mặt người trong ảnh I_i $(1 <= i <= N)$.
  - Một ảnh I_x chứa một khuôn mặt duy nhất cần được xác định danh tính.
- **Output:**
  - $\hat{y}_x$: Danh tính dự đoán của khuôn mặt trong ảnh I_x được xác định bằng cách chọn ảnh tham chiếu có độ tương đồng lớn nhất với I_x trong bộ dữ liệu tham chiếu. Nếu độ tương đồng lớn nhất này vượt ngưỡng cho trước, trả về danh tính của ảnh tham chiếu đó; ngược lại trả về "Unknown".

Hướng tiếp cận (1)

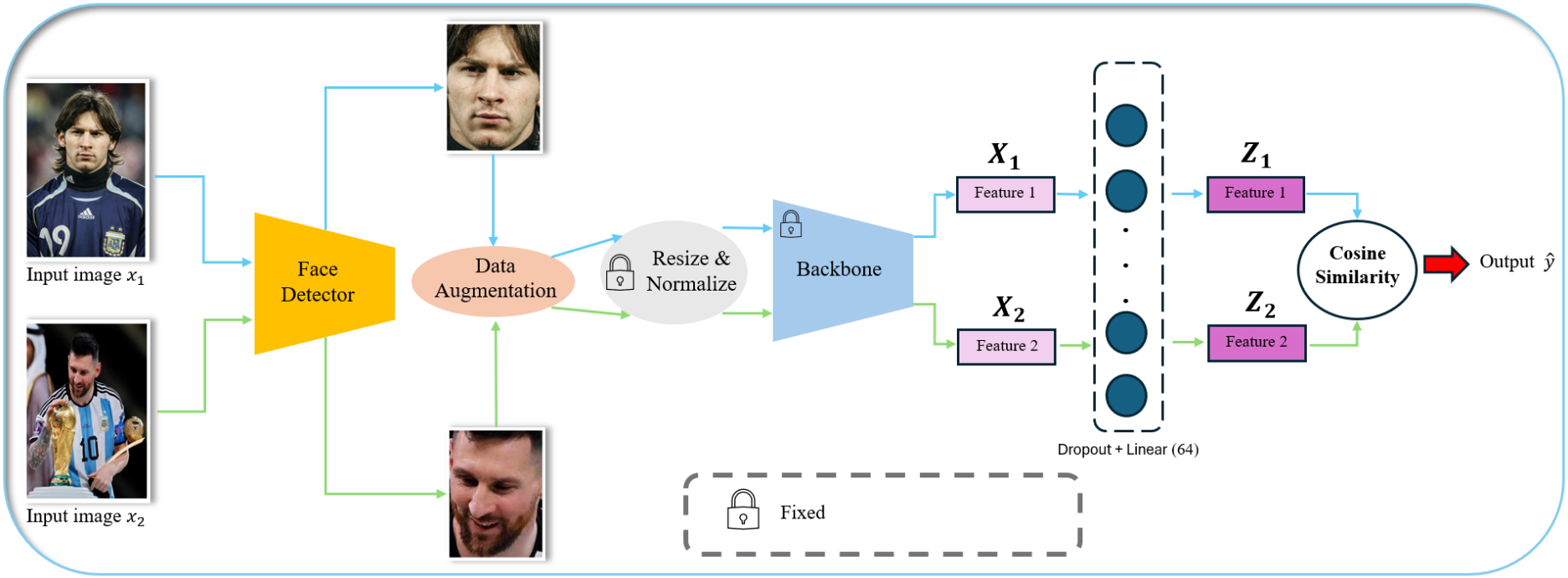
Ý tưởng: Tính xác suất hai ảnh đầu vào thuộc cùng một người



Hình 2: Pipeline hướng tiếp cận (1)

Hướng tiếp cận (2)

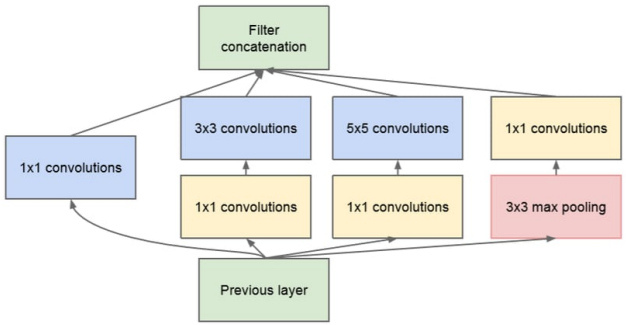
Ý tưởng: Tính mức độ tương đồng giữa hai ảnh đầu vào



Hình 3: Pipeline hướng tiếp cận (2)

Backbone - Inception-ResNet

- Inception-ResNet là một kiến trúc mạng kết hợp các đặc điểm nổi bật của InceptionNet và ResNet để cải thiện khả năng của mô hình.
- InceptionNet sử dụng song song nhiều kernel với kích thước khác nhau $(1 \times 1, 3 \times 3, 5 \times 5)$ rồi kết hợp đầu ra lại với nhau để tận dụng thông tin nhiều phạm vi không gian.

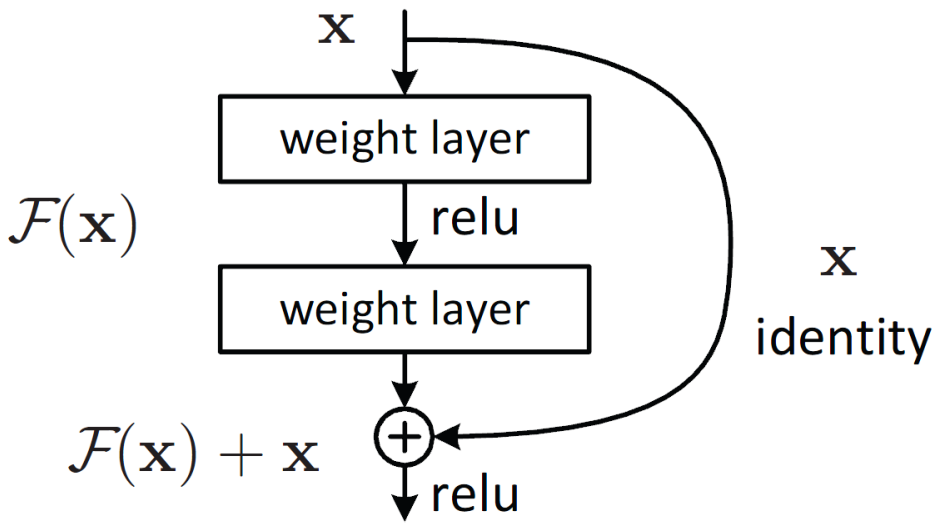


Hình 4: Inception module with dimension reductions

- Thêm các tầng tích chập 1x1 để giảm số lượng kênh trước khi thực hiện các phép tích chập lớn hơn, giúp giảm đáng kể số lượng tham số.

Backbone - Inception-ResNet

- ResNet giải quyết vấn đề vanishing gradient trong việc huấn luyện các mạng rất sâu bằng cách sử dụng thêm các kết nối phần dư (residual connection).



Hình 5: Residual Connection

⇒ Kiến trúc Inception-ResNet vừa duy trì khả năng xử lý linh hoạt của InceptionNet, vừa đảm bảo độ sâu và tính ổn định trong huấn luyện nhờ ResNet.

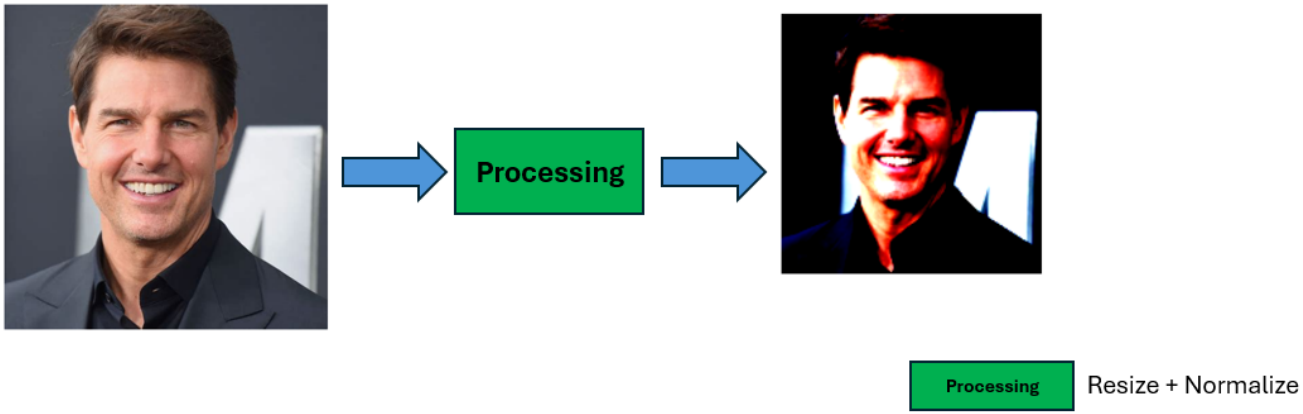
Face Detector - YOLOv5n

- Sử dụng YOLOv5n xác định chính xác vị trí khuôn mặt xuất hiện trong ảnh, giúp mô hình tập trung học các đặc trưng của khuôn mặt.
- YOLOv5n đã được nhóm fine-tuning trên bộ dữ liệu Face-Detection-Dataset (Kaggle).
  - Bộ dữ liệu này gồm 13400 mẫu dành cho huấn luyện và 3347 mẫu dùng để đánh giá.
  - Về cài đặt huấn luyện cho YOLOv5n: epoch = 30, batch size = 256 và các ảnh đều được resize về kích thước (640, 640).
  - Kết quả sau khi huấn luyện mAP50 = 0.875 đủ để thực hiện việc phát hiện khuôn mặt trong ảnh.

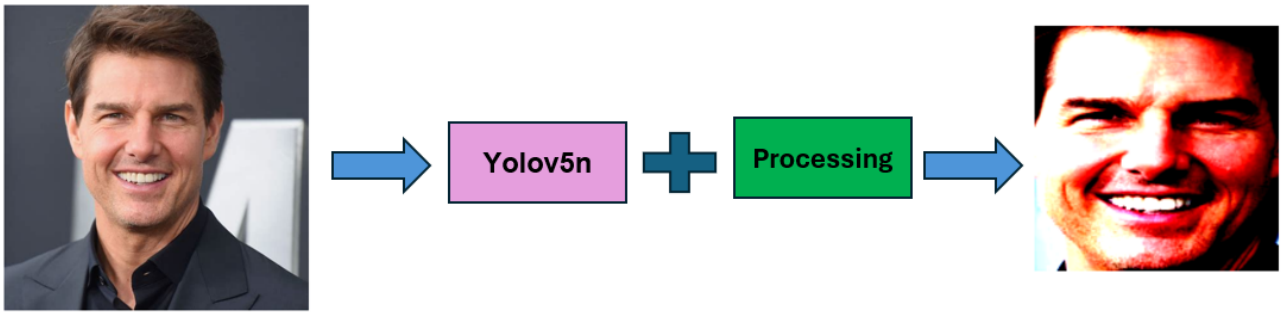
Image Pre-processing

- Ảnh đầu vào (hoặc ảnh được Face Detector xác định) được thực hiện Data Augmentation.
  - **Điều chỉnh độ sáng:** Ảnh sẽ được điều chỉnh độ sáng dựa vào việc sinh một giá trị ngẫu nhiên factor $\sim$ Uniform(0.5, 1.5). Nếu factor $= 1$ thì độ sáng ảnh giữ nguyên, factor > 1 thì ảnh sẽ sáng lên và ảnh sẽ tối đi nếu factor < 1.
  - **Lật ảnh:** Ảnh sẽ được lật theo chiều ngang với xác suất 0.5.
- Ảnh tiếp tục được resize về kích thước (160,160) và chuẩn hóa về phân phối chuẩn rồi mới qua Backbone.

Image Pre-processing



(a) Tiền xử lý ảnh cơ bản



(b) Mở rộng tiền xử lý ảnh

Hình 6: Một số phương pháp tiền xử lý sử dụng trong đồ án

LFW Dataset

- **Giới thiệu:** LFW (Labeled Faces in the Wild) là một bộ dữ liệu ảnh khuôn mặt, được sử dụng rộng rãi trong nghiên cứu về nhận diện khuôn mặt. Bộ dữ liệu này gồm các ảnh khuôn mặt của nhiều cá nhân nổi tiếng, được thu thập từ Internet.
- **Đặc điểm:**
  - Bộ dữ liệu chứa hơn 13,000 ảnh khuôn mặt, với 5,749 cá nhân.
  - Mỗi cá nhân có từ 1 đến 5 ảnh khuôn mặt, chụp trong các điều kiện khác nhau (góc nhìn, ánh sáng, biểu cảm).
  - LFW được sử dụng để đánh giá khả năng nhận diện khuôn mặt trong môi trường tự nhiên (không phải studio).
- **Mục đích sử dụng:** LFW chủ yếu được sử dụng để thử nghiệm các phương pháp nhận diện khuôn mặt, bao gồm xác thực khuôn mặt và phân biệt các cá nhân.

Celebrity Face Image Dataset

- Celebrity Face Image Dataset là một bộ dữ liệu được cung cấp bởi tác giả Vishesh Thakur trên Kaggle.
- Bộ dữ liệu chứa 1800 ảnh của 17 diễn viên Hollywood như: Leonardo DiCaprio, Johnny Depp, Tom Cruise, . . .



Hình 7: Ảnh trong bộ dữ liệu Celebrity Face Image

Accuracy

Accuracy: Đo lường tỷ lệ cặp ảnh được phân loại chính xác (cùng danh tính hoặc khác danh tính) trên tổng số cặp ảnh đầu vào.

$$Accuracy = \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)}{N} \tag{1}$$

Trong đó:

- N: Tổng số cặp ảnh đầu vào (x_{i1}, x_{i2}).
- $\hat{y}_i$: Nhãn dự đoán của mô hình cho cặp ảnh (x_{i1}, x_{i2}).

$$\hat{y}_i = \begin{cases} 1, & \text{Nếu } x_{i1} \text{ và } x_{i2} \text{ cùng danh tính} \\ 0, & \text{Nếu } x_{i1} \text{ và } x_{i2} \text{ khác danh tính} \end{cases} \tag{2}$$

- $\mathbb{1}(\hat{y}_i = y_i)$: Hàm chỉ thị, trả về 1 nếu $\hat{y}_i = y_i$, ngược lại trả về 0.

1-shot Accuracy

Với mỗi ảnh truy vấn q_i thì danh tính dự đoán của ảnh thỏa mãn:

$$\hat{y}_i = \arg \max_{c \in \{1,2,\ldots,C\}} simi(q_i, s_c) \tag{3}$$

1-shot Accuracy: Đo lường tỷ lệ dự đoán đúng của mô hình khi chỉ có một ảnh tham chiếu cho mỗi danh tính.

$$1 - shot\ Accuracy = \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)}{N} \tag{4}$$

Trong đó:

- $S = \{s_1, s_2, \ldots, s_C\}$: Tập các ảnh tham chiếu, mỗi ảnh s_c đại diện duy nhất cho danh tính c.
- $Q = \{q_1, q_2, \ldots, q_N\}$: Tập các ảnh cần truy vấn, gồm N ảnh.
- $simi(q_i, s_c)$: Hàm mức độ tương đồng giữa ảnh truy vấn q_i và ảnh tham chiếu s_c.

Cài đặt

Dưới đây là thông tin về cài đặt huấn luyện:

- Được huấn luyện với tổng số 100 epoch và áp dụng kỹ thuật Early Stopping với patience = 10.
- Ảnh trước khi vào Backbone sẽ được resize về kích thước (160,160) và sử dụng batch size = 256.
- Sử dụng thuật toán Adam với learning rate = 0.1 để tối ưu các tham số.
- Tập các ảnh tham chiếu được dùng để tính 1-shot Accuracy được chọn ngẫu nhiên.

Kết quả

Bảng 1: Accuracy trên bộ dữ liệu LFW (%)

Method		Dropout Rate	Data Augmentation	Accuracy
Hướng tiếp cận (1)	Inception-Resnet (v1)	0.75	No	92.17
	YOLOv5n + Inception-Resnet (v1)	0.5	No	**93.10**
Hướng tiếp cận (2)	YOLOv5n + Inception-Resnet (v1)	0.5	No	91.33
	YOLOv5n + Inception-Resnet (v1)	0.75	No	91.60

Bảng 2: 1-shot Accuracy trên bộ dữ liệu Celebrity Face Image Dataset

Method		Dropout Rate	Data Augmentation	1-shot Accuracy
Hướng tiếp cận (1)	Inception-Resnet (v1)	0.75	No	0.4375
	YOLOv5n + Inception-Resnet (v1)	0.5	No	**0.9798**
Hướng tiếp cận (2)	YOLOv5n + Inception-Resnet (v1)	0.5	No	0.9568
	YOLOv5n + Inception-Resnet (v1)	0.75	No	0.9209

Kết luận

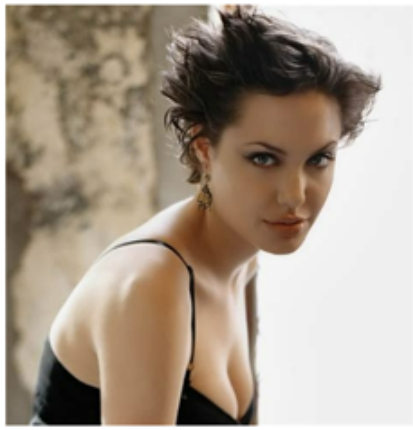
- **Nhận xét:**
  - Việc sử dụng YOLOv5n kết hợp với Inception-Resnet (v1) theo Hướng tiếp cận (1) đạt được kết quả cao nhất với 1-shot Accuracy=97.98% và Accuracy=93.10%.
  - Các phương pháp tăng cường dữ liệu được sử dụng chưa phát huy được hiệu quả.
- **Hạn chế:**
  - Chưa thử nghiệm với các bộ dữ liệu lớn và phức tạp hơn.
  - Hiệu suất vẫn phụ thuộc vào chất lượng phát hiện khuôn mặt bởi YOLO và quy trình tiền xử lý.

Định hướng phát triển

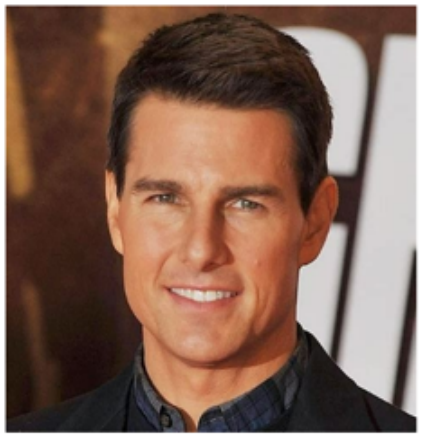
- **Tích hợp liveness detection:**
  - Nâng cao bảo mật bằng cách thêm tính năng phát hiện gian lận, giảm thiểu rủi ro từ các hình thức tấn công như sử dụng ảnh hoặc video giả mạo.
- **Cải tiến YOLO và pipeline nhận diện:**
  - Tối ưu hóa YOLO để nhận diện khuôn mặt trong các điều kiện ánh sáng, góc chụp và môi trường phức tạp.
  - Kết hợp các phương pháp mới trong pipeline xử lý để tăng độ chính xác.
- **Thử nghiệm trên các bộ dữ liệu lớn hơn:**
  - Mở rộng quy mô thí nghiệm trên các bộ dữ liệu đa dạng hơn, ví dụ MegaFace hoặc MS-Celeb-1M, để kiểm tra khả năng ứng dụng thực tế.
- **Ứng dụng thực tiễn:**
  - Triển khai thử nghiệm trong các hệ thống như chấm công, kiểm tra danh tính hoặc kiểm tra an ninh.

Demo

Johnny Depp (p ~= 0.9281)



Angelina Jolie (p ~= 0.9222)



Tom Cruise (p ~= 0.9903)

Megan Fox (p ~= 0.9175)

Hình 8: Minh họa Demo

Cảm ơn thầy và các bạn đã theo dõi!

Tài liệu tham khảo

- Going Deeper with Convolutions
- Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning
- Siamese Neural Networks for One-shot Image Recognition