**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**A logo with a fire and flames

Description automatically generated**

**ĐỒ ÁN CUỐI KỲ**

**THỊ GIÁC MÁY TÍNH**

**ĐỀ TÀI:**

**NHẬN DIỆN VÀ DỰ ĐOÁN GƯƠNG MẶT ĐEO KHẨU TRANG BẰNG THỊ GIÁC MÁY TÍNH**

Người thục hiện: **Trương Xuân Hùng – 21090141**

**Nguyễn Trọng Nghĩa – 21031271**

Lớp : **DHKHDL17A - 420300411601**

Khóa : **17**

Giảng viên hướng dẫn: **Vũ Đức Thịnh**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

**A logo with a fire and flames

Description automatically generated**

**ĐỒ ÁN CUỐI KỲ**

**THỊ GIÁC MÁY TÍNH**

**ĐỀ TÀI:**

**NHẬN DIỆN VÀ DỰ ĐOÁN GƯƠNG MẶT ĐEO KHẨU TRANG BẰNG THỊ GIÁC MÁY TÍNH**

Người thục hiện: **Trương Xuân Hùng – 21090141**

**Nguyễn Trọng Nghĩa – 21031271**

Lớp : **DHKHDL17A - 420300411601**

Khóa : **17**

Giảng viên hướng dẫn: **Vũ Đức Thịnh**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

# **ĐÁNH GIÁ CỦA GIẢNG VIÊN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |
| --- | --- |
|  | Tp. Hồ Chí Minh, ngày tháng năm  (kí và ghi họ tên) |

# **MỤC LỤC**

[ĐÁNH GIÁ CỦA GIẢNG VIÊN 1](#_Toc199285158)

[MỤC LỤC 2](#_Toc199285159)

[DANH MỤC CÁC HÌNH VẼ 4](#_Toc199285160)

[GIỚI THIỆU NỘI DUNG ĐỒ ÁN 5](#_Toc199285161)

[CHƯƠNG I. GIỚI THIỆU BÀI TOÁN 6](#_Toc199285162)

[1.1 Đặt vấn đề 6](#_Toc199285163)

[1.2 Ý nghĩa bài toán 6](#_Toc199285164)

[1.3 Thách thức của bài toán 6](#_Toc199285165)

[1.4 Phương pháp tiếp cận 7](#_Toc199285166)

[1.5 Cơ sở mạng nơ-ron tích chập CNN 7](#_Toc199285167)

[1.5.1 Giới Thiệu Về CNN 7](#_Toc199285168)

[1.5.2 Cách Hoạt Động Của CNN 8](#_Toc199285169)

[1.5.3 Trích Xuất Đặc Trưng và Phân Loại 9](#_Toc199285170)

[1.5.4 Tại Sao CNN Quan Trọng? 9](#_Toc199285171)

[1.5.5 Khi Nào Nên Sử Dụng CNN? 9](#_Toc199285172)

[1.6 Cơ Sở Lý Thuyết Về YOLO (You Only Look Once) 9](#_Toc199285173)

[1.6.1 Giới Thiệu 9](#_Toc199285174)

[1.6.2 Kiến Trúc Mô Hình 10](#_Toc199285175)

[1.6.3 Nguyên Lý Hoạt Động 10](#_Toc199285176)

[1.6.4 Ưu Điểm Của YOLO 12](#_Toc199285177)

[1.6.5 Nhược Điểm Của YOLO 12](#_Toc199285178)

[1.6.6 Các Phiên Bản Của YOLO 12](#_Toc199285179)

[1.6.7 Ứng Dụng 12](#_Toc199285180)

[1.7 Cơ Sở Lý Thuyết Về U-Net 13](#_Toc199285181)

[1.7.1 Giới thiệu 13](#_Toc199285182)

[1.7.2 Kiến trúc mô hình 13](#_Toc199285183)

[1.7.3 Nguyên lý hoạt động 14](#_Toc199285184)

[CHƯƠNG II. THU THẬP DỮ LIỆU 15](#_Toc199285185)

[2.1 Thu thập dữ liệu 15](#_Toc199285186)

[2.2 Dữ liệu 15](#_Toc199285187)

[2.3 Tiền xử lý dữ liệu 17](#_Toc199285188)

[CHƯƠNG III. XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH 20](#_Toc199285189)

[3.1 Xây dựng mô hình 20](#_Toc199285190)

[3.2 Huấn luyện mô hình 20](#_Toc199285191)

[CHƯƠNG IV. ĐÁNH GIÁ KẾT QUẢ MÔ HÌNH 22](#_Toc199285192)

[4.1 Đánh giá kết quả mô hình YOLO 22](#_Toc199285193)

[4.1.1 Đánh giá loss của epoch cuối cùng 22](#_Toc199285194)

[4.1.2 Đánh giá trên tập Validation và Test 22](#_Toc199285195)

[4.2 Đánh giá kết quả mô hình U-Net 24](#_Toc199285196)

[4.2.1 Thang đo SSIM, PSNR, LPIPS và MAE 24](#_Toc199285197)

[4.2.2 Hình ảnh đầu ra 25](#_Toc199285198)

[CHƯƠNG V. TỔNG KẾT 26](#_Toc199285199)

[TÀI LIỆU THAM KHẢO 27](#_Toc199285200)

# **DANH MỤC CÁC HÌNH VẼ**

[Hình 1.5.1 Sơ đồ mạng Nơ-ron 10](#_Toc199323307)

[Hình 1.5.2 Sơ đồ mạng nơ-ron tích chập (CNN) 10](#_Toc199323308)

[Hình 1.6.2: Sơ đồ kiến trúc mạng YOLO 12](#_Toc199323309)

[Hình 1.6.3 Kiến trúc của 1 output model YOLO 13](#_Toc199323310)

[Hình 1.7.3 Sơ đồ hoạt động của U-Net 16](#_Toc199323311)

[Hình 2.3.1. Minh họa hình ảnh trong bộ dữ liệu 18](#_Toc199323312)

[Hình 2.2.2. Phân phối số lượng hình ảnh theo nhóm "có khẩu trang" và "không khẩu trang" 18](#_Toc199323313)

[Hình 2.3. Ảnh gương mặt sau khi thêm khẩu trang bằng MaskTheFace 19](#_Toc199323314)

[Hình 2.3.1 Dữ liệu trước và sau khi được dán nhãn 20](#_Toc199323315)

[Hình 2.3.3 Phân bố điểm ảnh trước và sau khi chuẩn hóa về khoảng [0,1] 21](#_Toc199323316)

[Hình 4.1.2. Biểu đồ kết quả huấn luyện 26](#_Toc199323317)

[Hình 4.1.3 Hình ảnh dự đoán của mô hình sau khi train 26](#_Toc199323318)

[Hình 4.2.1 Biểu đồ kết quả huấn luyện với SSIM và PSNR 27](#_Toc199323319)

[Hình 4.2.1.2 Biểu đồ kết quả huấn luyện với MAE Loss, MAE, MSE 27](#_Toc199323320)

[Hình 4.2.2 Kết quả ảnh đầu ra của U-Net 28](#_Toc199323321)

# **GIỚI THIỆU NỘI DUNG ĐỒ ÁN**

Nhận diện gương mặt là một bài toán quan trọng trong lĩnh vực thị giác máy tính, với nhiều ứng dụng thực tiễn trong an ninh, giám sát, y tế và kiểm soát dịch bệnh. Đặc biệt, trong bối cảnh đại dịch COVID-19, nhu cầu phát hiện và xác định xem một người có đeo khẩu trang hay không trở nên cấp thiết để đảm bảo tuân thủ các quy định y tế và bảo vệ an toàn cộng đồng.

Trong đồ án này, nhóm tập trung phát triển và phân tích hiệu suất của hai mô hình học sâu: mô hình YOLO (You Only Look Once) để nhận dạng gương mặt đeo khẩu trang và mô hình U-Net để khôi phục hình ảnh gương mặt bị che khuất bởi khẩu trang. YOLO được sử dụng để phát hiện nhanh và chính xác các vùng gương mặt, xác định xem gương mặt có đeo khẩu trang hay không. Trong khi đó, U-Net được áp dụng để tái tạo lại các đặc điểm gương mặt bị che khuất, giúp khôi phục hình ảnh gương mặt một cách chi tiết và chân thực.

Mục tiêu của đồ án là đánh giá độ chính xác, tốc độ xử lý và khả năng tổng quát hóa của hai mô hình này trên tập dữ liệu gồm 10.000 ảnh gương mặt, bao gồm cả ảnh có và không có khẩu trang. Bộ dữ liệu được chia thành ba phần: tập huấn luyện, tập kiểm tra và tập xác thực. Nhóm tiến hành các thí nghiệm để so sánh hiệu suất của YOLO và U-Net, từ đó đề xuất giải pháp tối ưu cho bài toán nhận diện và khôi phục gương mặt hoặc có thể xây dựng một hệ thống có thể vừa nhận diện và vừa tái tạo lại gương mặt khi đeo khẩu trang.

# **CHƯƠNG I. GIỚI THIỆU BÀI TOÁN**

* 1. **Đặt vấn đề**

Trong thời đại công nghệ số và trí tuệ nhân tạo (AI) phát triển mạnh mẽ, các hệ thống nhận diện khuôn mặt ngày càng được ứng dụng rộng rãi trong nhiều lĩnh vực như an ninh, kiểm soát truy cập, giám sát thông minh và chăm sóc sức khỏe. Đặc biệt, sau đại dịch COVID-19, việc đeo khẩu trang đã trở thành một biện pháp quan trọng để phòng chống dịch bệnh và đảm bảo an toàn nơi công cộng.

Tuy nhiên, khẩu trang che phủ một phần lớn khuôn mặt, gây khó khăn cho các hệ thống nhận diện khuôn mặt truyền thống. Do đó, việc phát triển một hệ thống có khả năng tự động nhận diện và phân loại người có đeo khẩu trang hay không đeo khẩu trang trở nên cần thiết, giúp nâng cao hiệu quả kiểm soát y tế, an toàn tại các địa điểm công cộng như sân bay, trung tâm thương mại, trường học và bệnh viện.

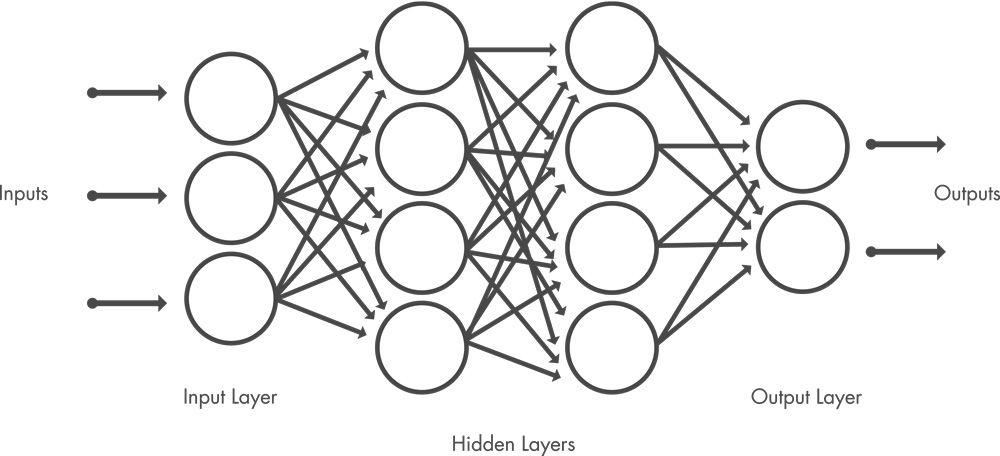
Bài toán này thuộc lĩnh vực thị giác máy tính (Computer Vision) và có thể được giải quyết bằng các phương pháp học sâu (Deep Learning). Trong nghiên cứu này, nhóm thực hiện phân tích và đánh giá hiệu suất của phương pháp YOLO (You Only Look Once) cho phân tích dự đoán khẩu trang và U-Net cho phân tích dự đoán gương mặt.

* 1. **Ý nghĩa bài toán**
* Đối với lĩnh vực y tế và an toàn công cộng: Hỗ trợ kiểm soát việc tuân thủ quy định đeo khẩu trang, đặc biệt tại các khu vực đông người, giúp giảm nguy cơ lây lan dịch bệnh.
* Đối với các hệ thống giám sát an ninh: Kết hợp với nhận diện khuôn mặt để nâng cao khả năng xác thực danh tính ngay cả khi người dùng đeo khẩu trang.
* Đối với tự động hóa và AI: Đóng góp vào sự phát triển của các ứng dụng thị giác máy tính, đặc biệt trong lĩnh vực nhận diện đối tượng và phân loại hình ảnh.
* Đối với nghiên cứu khoa học: Cung cấp cơ sở dữ liệu và phương pháp nghiên cứu mới, giúp cải thiện độ chính xác của các mô hình nhận diện khuôn mặt trong điều kiện che khuất một phần.
  1. **Thách thức của bài toán**
* Độ che phủ của khẩu trang: Khẩu trang che đi phần lớn các đặc điểm quan trọng trên khuôn mặt, gây khó khăn cho các mô hình nhận diện truyền thống.
* Sự đa dạng của khẩu trang: Màu sắc, kích thước, kiểu dáng và chất liệu của khẩu trang có thể ảnh hưởng đến độ chính xác của mô hình.
* Sự đa dạng của khuôn mặt: Sự khác biệt về giới tính, độ tuổi, màu da, hình dạng khuôn mặt và các yếu tố môi trường (ánh sáng, góc nhìn, nền ảnh) có thể làm giảm hiệu suất của hệ thống nhận diện.
* Cân bằng dữ liệu: Đảm bảo tập dữ liệu huấn luyện có sự cân bằng giữa các lớp (đeo khẩu trang và không đeo khẩu trang) để tránh hiện tượng overfitting.
* Tốc độ xử lý và triển khai thực tế: Đối với các ứng dụng giám sát thời gian thực, mô hình phải có tốc độ xử lý nhanh mà vẫn đảm bảo độ chính xác cao.
  1. **Phương pháp tiếp cận**

Để giải quyết bài toán nhận diện đeo khẩu trang, nhóm sử dụng phương pháp chính trong học sâu:

* **CNN (Convolutional Neural Network):** CNN là một trong những mô hình mạnh mẽ trong nhận diện hình ảnh. Mô hình được huấn luyện trên tập dữ liệu hình ảnh khuôn mặt có và không đeo khẩu trang, giúp trích xuất các đặc trưng quan trọng để phân loại
* **YOLO (You Only Look Once):** YOLO là một mô hình nhận diện đối tượng thời gian thực, có khả năng xác định và khoanh vùng khuôn mặt, đồng thời phân loại xem người đó có đeo khẩu trang hay không. Phương pháp này phù hợp với các hệ thống giám sát tự động.
* **U-NET:** U-Net làmột kiến trúc mạng nơ-ron tích chập (CNN) được thiết kế đặc biệt cho các tác vụ phân đoạn ảnh (image segmentation), nhưng nó cũng rất hiệu quả trong nhiều bài toán tái tạo và phục hồi ảnh, bao gồm cả inpainting khuôn mặt.
  1. **Cơ sở mạng nơ-ron tích chập CNN**
     1. **Giới Thiệu Về CNN**

Mạng nơ-ron tích chập (Convolutional Neural Network - CNN) là một loại mô hình học sâu được thiết kế để xử lý dữ liệu dạng hình ảnh và chuỗi thời gian. CNN giúp tự động học và phát hiện các đặc trưng quan trọng trong hình ảnh mà không cần tiến hành xử lý dữ liệu một cách thủ công.

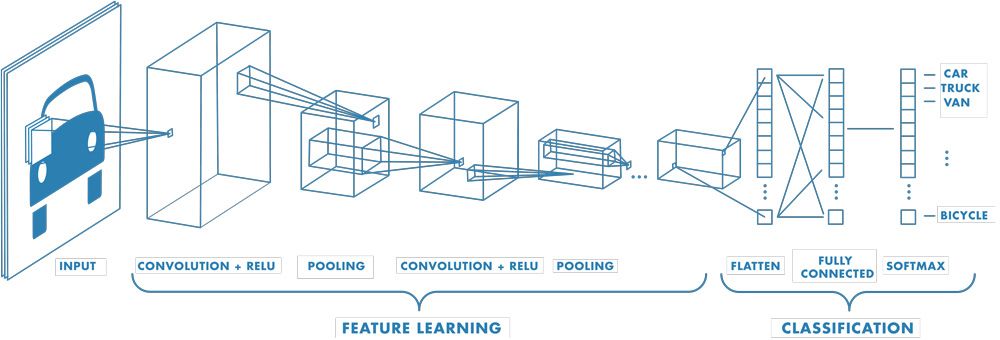


Hình 1.5.1 Sơ đồ mạng Nơ-ron

* + 1. **Cách Hoạt Động Của CNN**

Một mô hình CNN bao gồm ba loại lớp chính:

* Lớp tích chập (Convolutional Layer)
* Lớp phi tuyến tính (ReLU - Rectified Linear Unit)
* Lớp lược bỏ (Pooling Layer)



Hình 1.5.2 Sơ đồ mạng nơ-ron tích chập (CNN)

1. **Lớp Tích Chập (Convolutional Layer)**

* Áp dụng các bộ lọc (filters/kernel) trên hình ảnh để tạo ra các đặc trưng quan trọng.
* Các bộ lọc ban đầu phát hiện những đặc trưng đơn giản như cạnh, góc, màu sắc, v.v.
* Càng nhiều lớp tích chập, mô hình càng học được những đặc trưng phức tạp hơn.

1. **Lớp Phi Tuyến Tính (ReLU - Rectified Linear Unit)**

* Giúp tăng tốc độ huấn luyện bằng cách loại bỏ giá trị âm (chính là giá trị âm trở thành 0).
* Đảm bảo mối lớp tích chập chỉ truyền đi những đặc trưng quan trọng nhất.

1. **Lớp Lược Bỏ (Pooling Layer)**

Giúm giảm kích thước dữ liệu, tăng tốc độ tính toán.

Hai loại phổ biến:

* Max Pooling: Chọn giá trị lớn nhất trong một vùng.
* Average Pooling: Tính trung bình giá trị trong một vùng.
  + 1. **Trích Xuất Đặc Trưng và Phân Loại**

Sau khi học các đặc trưng qua các lớp tích chập và lược bỏ, CNN chuyển sang giai đoạn phân loại:

* Lớp Fully Connected: Kết nối tất cả các đặc trưng đã học và đổ vào một vector K chiều (với K là số lượng lớp cần phân loại).
* Lớp Softmax: Chuyển đổ vector Fully Connected thành xác suất của từng lớp.
  + 1. **Tại Sao CNN Quan Trọng?**

CNN được ứng dụng rộng rãi trong nhiều lĩnh vực:

* Y học: Phát hiện ung thư từ ảnh chụp X-ray, MRI.
* Xử lý âm thanh: Nhận diện giọng nói, tìm kiếm từ khoá.
* Phát hiện đối tượng: Tích hợp trong xe tự hành, nhận diện khuôn mặt.
* Tạo dữ liệu giả: GANs tạo ảnh giả để huấn luyện mô hình AI.
  + 1. **Khi Nào Nên Sử Dụng CNN?**
* Khi làm việc với dữ liệu phức tạp như hình ảnh, video.
* Khi xử lý tín hiệu thời gian như chuỗi giọng nói.
* Khi muốn sử dụng các mô hình tiên tiến như AlexNet, GoogLeNet, YOLO.

Với những đặc điểm trên, CNN là lựa chọn tối ưu cho nhiều bài toán xử lý hình ảnh và nhận dạng mẫu.

* 1. **Cơ Sở Lý Thuyết Về YOLO (You Only Look Once)**
     1. **Giới Thiệu**

YOLO (You Only Look Once) là một trong những kiểu mô hình nhận dạng và phân loại vật thể trong ảnh và video nhanh và hiệu quả nhất. Khác với các mô hình truyền thống như R-CNN hay Fast R-CNN, YOLO xác định vị trí và phân loại vật thể trong đối tượng chỉ trong một lượt xem duy nhất, giúp giảm đáng kể thời gian xử lý.

* + 1. **Kiến Trúc Mô Hình**

Mô hình YOLO sử dụng mạng nơ-ron tích chập (CNN) để xử lý hình ảnh đầu vào. Hình ảnh được chia thành một lưới, mỗi ô trong lưới dự đoán một số hộp giới hạn (bounding box) và xác suất đối tượng thuộc vào các lớp nhất định. Cách tiếp cận này giúp YOLO đạt được tốc độ nhanh mà vẫn đảm bảo tính chính xác.

A diagram of a diagram of a box

AI-generated content may be incorrect.

Hình 1.6.2: Sơ đồ kiến trúc mạng YOLO

Mô hình YOLO gồm các tầng chính:

* Thành phần Darknet Architechture được gọi là base network có tác dụng trích suất đặc trưng.
* Tầng convolutional: Xử lý để phát hiện đặc trưng.
* Tầng fully connected: Biến đổi dữ liệu để xác định bounding boxes.
  + 1. **Nguyên Lý Hoạt Động**

YOLO sẽ chia ảnh thành một lưới vuông (có dạng như ma trận SxS). Mỗi ô trong lưới sẽ dự đoán: N hộp giới hạn (bounding boxes), độ tin cậy (confidence score), và các nhãn vật thể

Các bức ảnh khi được đưa vào mô hình sẽ được scale để về chung một kích thước phù hợp với input shape của mô hình và sau đó được gom lại thành batch đưa vào huấn luyện**.**

Sau khi đi qua các layer convolutional thì shape giảm dần theo cấp số nhân là 2. Cuối cùng ta thu được một feature map có kích thước tương đối nhỏ để dự báo vật thể trên từng ô của feature map.

Output của mô hình YOLO là một véc tơ sẽ bao gồm các thành phần:

A black text on a white background

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Hình 1.6.3 Kiến trúc của 1 output model YOLO

Tóm tắt hơn thì YOLO sẽ hoạt động như sau

1. Hệ thống nhận dạng ảnh và chia thành lưới.
2. Mỗi ô trong lưới sẽ đề xuất N bounding boxes cùng điểm tin cậy.
3. Giá trị IOU (Intersection over Union) sẽ được sử dụng để loại các hộp giới hạn không chính xác.
4. Thuật toán NMS (Non-Maximum Suppression) giúp giảm bớt các bounding boxes không có giá trị.
   * 1. **Ưu Điểm Của YOLO**

* Tốc độ cao: YOLO nhanh hơn các mô hình truyền thống nhờ xử lý vật thể trong một lượt xem.
* Dự đoán toàn cục: Nhận dạng toàn bối cảnh, giúp hạn chế sai sót khi xác định vật thể.
* Tính tổng quát tốt: Có thể áp dụng vào nhiều loại ảnh khác nhau.
  + 1. **Nhược Điểm Của YOLO**
* Khó xác định vật thể nhỏ: Do chía ảnh thành lưới, các vật thể nhỏ có thể bị bỏ lỡ.
* Giá trị IOU cao: Không hoàn hảo khi xử lý các vật thể chồng chéo.
  + 1. **Các Phiên Bản Của YOLO**
* **YOLOv1**: Phiên bản đầu tiên, giới thiệu cách tiếp cận mới nhưng vẫn còn hạn chế về độ chính xác.
* **YOLOv2 (YOLO9000):** Cải tiến độ chính xác và khả năng nhận diện nhiều lớp đối tượng hơn.
* **YOLOv3:** Sử dụng kiến trúc Darknet-53 để nâng cao hiệu suất.
* **YOLOv4:** Tối ưu hóa tốc độ và độ chính xác, phù hợp với nhiều ứng dụng thực tế.
* **YOLOv5:** Do Ultralytics phát triển, cải thiện về kích thước mô hình và dễ dàng triển khai.
* **YOLOv6, YOLOv7, YOLOv8:** Các phiên bản mới nhất với những cải tiến vượt bậc về hiệu suất và tốc độ.
  + 1. **Ứng Dụng**

YOLO được ứng dụng rộng rãi trong nhiều lĩnh vực:

* Giám sát an ninh: Nhận diện khuôn mặt, phát hiện hành vi bất thường trong hệ thống camera an ninh.
* Y tế: Phát hiện tổn thương trên ảnh X-quang hoặc MRI.
* Giao thông: Nhận diện phương tiện, phát hiện biển báo giao thông trong các hệ thống xe tự hành.
* Nông nghiệp: Giám sát cây trồng, phát hiện sâu bệnh trên hình ảnh thu được từ drone.
  1. **Cơ Sở Lý Thuyết Về U-Net**
     1. **Giới thiệu**

U-Net là một kiến trúc mạng nơ-ron tích chập (CNN) ban đầu U-Net được phát triển để phân đoạn phục vụ cho phân đoạn ảnh y tế, nhưng sau này nó trở nên hữu ích cho nhiều tác vụ khác, chẳng hạn như tô màu hình ảnh hoặc tô màu hình ảnh. Điểm mạnh của U-Net là có thể học tốt từ một tập dữ liệu nhỏ và cho ra kết quả phân đoạn ảnh chính xác, rõ ràng ở cấp độ pixel.

* + 1. **Kiến trúc mô hình**

**a) Contracting Path (Encoder)**

- Mục tiêu là để trích xuất đặc trưng và giảm dần kích thước ảnh (downsampling)

- Gồm nhiều block có cấu trúc:

+2 lớp Convolution 3x3 (padding = same) + ReLU

+1 lớp MaxPooling 2x2 (giảm kích thước ảnh xuống còn 1/2)

- Sau mỗi block, số lượng kênh (feature maps) tăng gấp đôi để lưu giữ nhiều đặc trưng hơn.

**b)** **Expanding Path (Decoder)**

- Mục tiêu là tăng dần kích thước ảnh (upsampling) để trở lại kích thước gốc. Gồm nhiều block có cấu trúc:

+ 1 lớp UpSampling 2x2 (hoặc ConvTranspose2D)

+ Ghép nối (concatenate) với feature maps tương ứng từ Encoder thông qua skip connections

+ 2 lớp Convolution 3x3 + ReLU

**c) Skip Connections**

- Mỗi tầng của Encoder được nối trực tiếp sang tầng tương ứng trong Decoder.

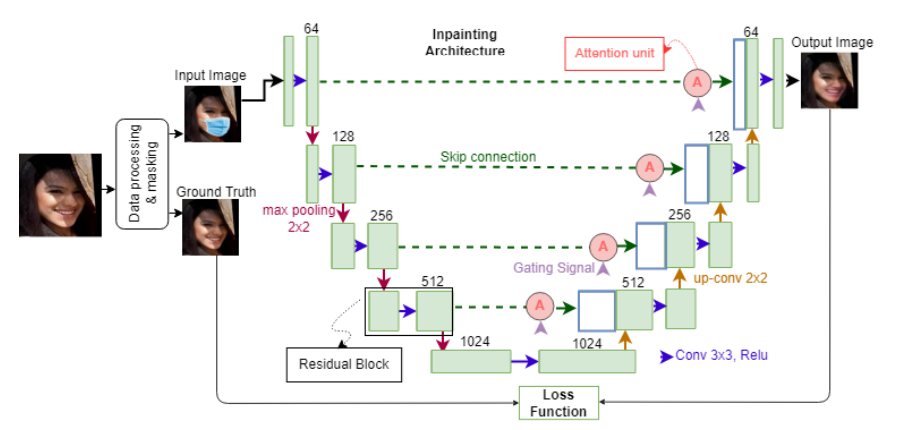
- Tác dụng:

+ Giữ lại thông tin không gian quan trọng mà các lớp downsampling có thể làm mất.

+ Giúp mô hình phục hồi chi tiết tốt hơn ở ảnh đầu ra.

**d) Output Layer**

* Lớp cuối cùng là **1 Convolution 1x1**, đưa số kênh về **số lớp phân đoạn (classes)**.
* Với bài toán phân đoạn nhị phân: đầu ra là một kênh (1-channel) chứa xác suất mỗi pixel thuộc lớp foreground (đối tượng).
  + 1. **Nguyên lý hoạt động**
* **Đầu vào ảnh** được đưa vào mô hình.
* **Encoder** trích xuất đặc trưng dần theo từng mức, đồng thời giảm độ phân giải để thu hẹp vùng nhìn (receptive field).
* **Decoder** sử dụng các thông tin đặc trưng trừu tượng từ encoder và kết hợp với thông tin chi tiết từ skip connections để tái tạo ảnh.
* **Ảnh đầu ra** có cùng kích thước với ảnh đầu vào, nhưng mỗi pixel thể hiện **xác suất thuộc lớp (**với softmax/sigmoid activation).

****

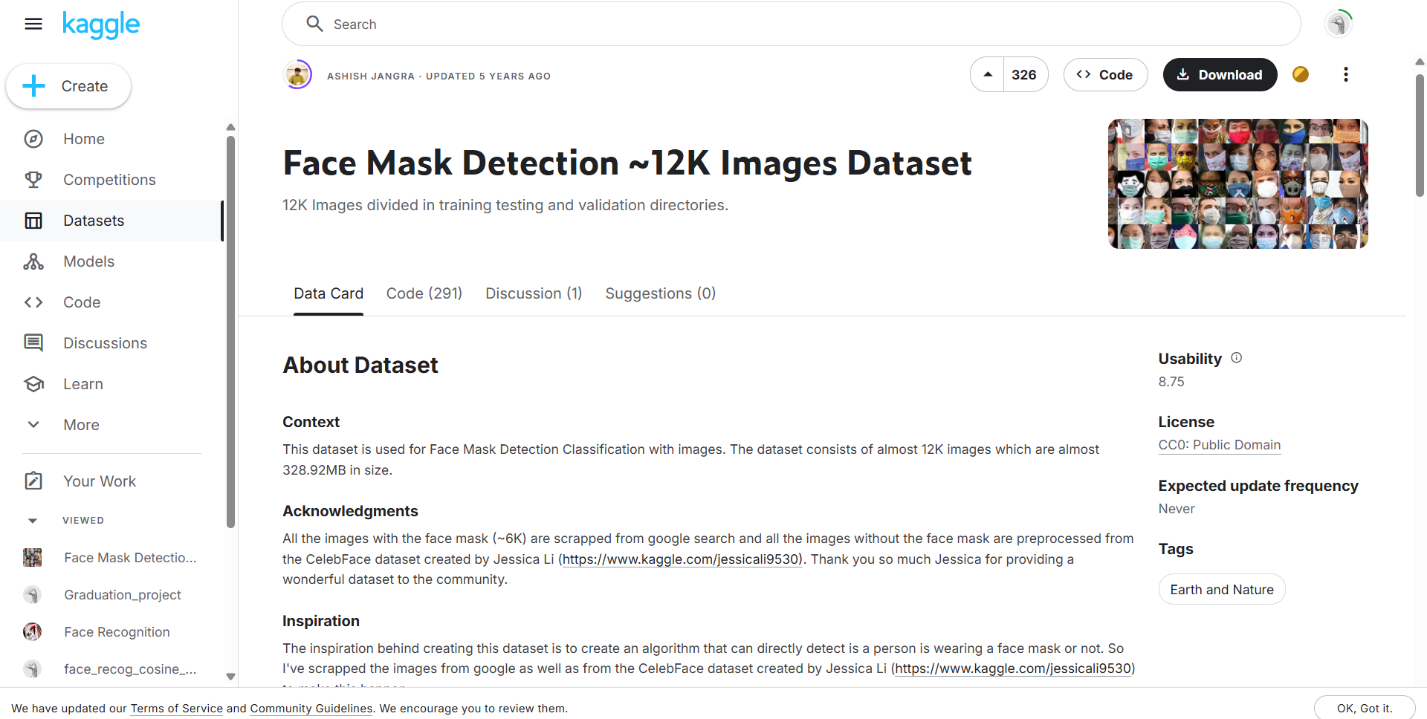
Hình 1.7.3 Sơ đồ hoạt động của U-Net

# **CHƯƠNG II. THU THẬP DỮ LIỆU**

**2.1 Thu thập dữ liệu**

**﻿** Để có thể tiến hành việc huấn luyện, xây dựng và đánh giá mô hình nhận diện khẩu trang, cần có một bộ dữ liệu đủ lớn, chứa đủ các trường hợp đeo và không đeo khẩu trang để huấn luyện mô hình nhận diện.

Thu thập dữ liệu từ Kaggle Trong nghiên cứu này, bộ dữ liệu được sử dụng là "Face Mask Detection Dataset", được thu thập từ Kaggle (https://www.kaggle.com/datasets/ashishjangra27/face-mask-12k-images-dataset).



Bộ dữ liệu này có khoảng 12.000 hình ảnh khuôn mặt, trong đó được chia thành hai nhóm:

* Có khẩu trang: Hình ảnh khuôn mặt người đeo khẩu trang.
* Không khẩu trang: Hình ảnh khuôn mặt người không đeo khẩu trang.

Bộ dữ liệu hình ảnh có các điều kiện khác nhau về góc nhìn, ánh sáng, biểu cảm khuôn mặt, giới tính và độ tuổi, giúp mô hình học được các đặc trưng đa dạng.

**2.2 Dữ liệu**

Bộ dữ liệu bao gồm 12.000 hình ảnh, chia thành hai nhóm có khẩu trang và không đeo khẩu trang.

A collage of two people wearing a mask

AI-generated content may be incorrect.

Hình 2.3.1. Minh họa hình ảnh trong bộ dữ liệu

*A red and green circle with text

AI-generated content may be incorrect.*

*A graph of a number of datasets

AI-generated content may be incorrect.*

Hình 2.2.2. Phân phối số lượng hình ảnh theo nhóm "có khẩu trang" và "không khẩu trang"

Biểu đồ phân phối cho thấy hai nhóm dữ liệu tương đối cân bằng, giúp giảm thiểu bias trong mô hình huấn luyện. Dữ liệu bao gồm các điều kiện ánh sáng và góc nhìn khác nhau, đảm bảo tính đa dạng khi huấn luyện.

## **2.3 Tiền xử lý dữ liệu**

Xử lý dữ liệu là quá trình chuẩn bị dữ liệu trước khi đưa vào huấn luyện mô hình nhận diện khẩu trang. Xử lý dữ liệu là quá trình thao tác và biến đổi dữ liệu từ các nguồn khác nhau thành dạng có ý nghĩa và có thể sử dụng được cho việc phân tích, đánh giá, hoặc sử dụng cho các mô hình dữ liệu.

Với bộ dữ liệu trên nhóm chỉ sử dụng phần bộ dữ liệu gương mặt không đeo khẩu trang (Without Mask). Vì để huấn luyện mô hình U-Net, cần sử dụng cặp ảnh mặt người có đeo khẩu trang và không đeo khẩu trang tương tự nhau. Nhóm đã sử dụng một project có sẵn trên Github (https://github.com/aqeelanwar/MaskTheFace.git) để thực hiện việc thêm khẩu trang cho những ảnh gương mặt chưa có khẩu trang. Việc này giúp tạo một cặp dữ liệu ảnh (ảnh gương mặt và ảnh gương mặt có khẩu trang) phù hợp cho 2 mô hình YOLO và U-Net.

A collage of people wearing masks

AI-generated content may be incorrect.

Hình 2.3. Ảnh gương mặt sau khi thêm khẩu trang bằng MaskTheFace

Quá trình xử lý dữ liệu bao gồm một loạt các bước nhằm đảm bảo dữ liệu đủ tính chính xác và từ đó có thể sử dụng hiệu quả. Các bước xử lý bao gồm:

* Xóa dữ liệu rỗng: Loại bỏ các hình ảnh không hợp lệ hoặc bị lỗi, thực tế lượng ảnh sử dụng cho cả 2 mô hình YOLO và U-Net sau khi xóa ảnh không hợp lệ còn khoảng 4450 ảnh.
* Chuẩn hóa kích thước: Mọi hình ảnh được điều chỉnh về cùng kích thước (256x256 pixel) để đảm bảo sự nhất quán khi đưa vào mô hình.
* Dán nhãn dữ liệu: Bộ dữ liệu hình ảnh vẫn chưa được dán nhãn nên nhóm đã sử dụng phần mềm LabelImg để dán nhãn thủ công cho từng bức ảnh.

A screenshot of a computer screen

AI-generated content may be incorrect.

Hình 2.3.1 Dữ liệu trước và sau khi được dán nhãn

* Tách bộ dữ liệu: Bộ dữ liệu đã được tác giả phân thành 2 nhóm đeo khẩu trang (WithMask) và không đeo khẩu trang (WithoutMask) sẵn.
* Chuẩn hóa dữ liệu: Với việc sử dụng các mô hình như CNN hoặc Pre-trained thì chuẩn hóa là một việc vô cùng quân trọng. Dữ liệu sẽ được tách thành 3 tập dữ liệu riêng biệt dành cho việc huấn luyện, kiểm thử và xác thực. Dữ liệu được chuẩn hóa về khoảng [0,1] bằng phương pháp Min-Max Scaling để giúp mô hình học tốt hơn. Công thức chuẩn hóa như sau:

Với là tập huấn luyện

A comparison of a graph

AI-generated content may be incorrect.

Hình 2.3.3 Phân bố điểm ảnh trước và sau khi chuẩn hóa về khoảng [0,1]

**CHƯƠNG III. XÂY DỰNG VÀ HUẤN LUYỆN MÔ HÌNH**

**3.1 Xây dựng mô hình**

**YOLO:**

Mô hình được xây dựng dựa trên kiến trúc YOLOv8 nano (yolov8n.pt) , một mạng nơ-ron tích chập (CNN) tối ưu cho bài toán phát hiện đối tượng, với trọng số được huấn luyện trước trên tập dữ liệu COCO để tăng tốc hội tụ. Mô hình YOLOv8 nano có 3,006,038 triệu tham số và nhóm dùng optimizer AdamW với 72 lớp cho mô hình nhận diện này.

Mô hình được thiết kế nhẹ, phù hợp với số lượng tài nguyên hạn chế, và sử dụng file cấu hình data.yaml để định nghĩa tập dữ liệu với lớp khuôn mặt và lớp khẩu trang.

**U-Net:**

Trong nghiên cứu này, nhóm đã xây dựng mô hình Residual Attention U-Net nhằm phục hồi khuôn mặt không bị che từ ảnh có khẩu trang. Các khối Attention được tích hợp tại các kết nối skip giữa encoder và decoder để mô hình có thể tập trung học tốt hơn ở các vùng quan trọng bị che như mũi và miệng. Ngoài ra, các kết nối residual giúp duy trì thông tin và giảm độ sâu hiệu dụng trong khi vẫn giữ được khả năng biểu diễn. Mô hình có 2,192,483 triệu tham số với 72 lớp và dùng optimizer Adam.

Dữ liệu được tổ chức trong hai thư mục: raw\_images chứa ảnh gốc không khẩu trang và raw\_images\_masked chứa các ảnh đã bị che bằng khẩu trang tổng hợp bằng công cụ MaskTheFace. Các cặp ảnh được ghép theo tên tệp để làm dữ liệu đầu vào và nhãn tương ứng. Mô hình có cấu trúc dựa trên U-Net với năm tầng encoder và năm tầng decoder, mỗi tầng gồm hai lớp convolution kèm theo Batch Normalization và ReLU.

## **3.2 Huấn luyện mô hình**

Hai mô hình sẽ được huấn luyện với dữ liệu huấn luyện và xác thực bằng dữ liệu xác thực và cuối cùng là kiểm thử lại với bộ dữ liệu kiểm thử.

**YOLO**

Mô hình được huấn luyện trong 100 epoch với batch size 32, sử dụng optimizer AdamW với learning rate ban đầu 0.01, giảm dần đến 0.001, và áp dụng early stopping sau 20 epoch không cải thiện. Tập dữ liệu huấn luyện được chia thành ba phần với 70% cho tập huấn luyện, 15% cho tập kiểm thử và 15% cho tập xác thực. Kỹ thuật tăng cường dữ liệu như mosaic, lật ảnh, xoay, và điều chỉnh HSV được sử dụng để tăng tính khái quát hóa. Hàm loss tổng hợp gồm CIoU Loss (box), Binary Cross-Entropy (classification), và Distribution Focal Loss (Dfl) được tối ưu trong quá trình huấn luyện. Hiệu suất được đánh giá trên tập validation và test bằng các thang đo mAP@0.5, mAP@0.5:0.95, Precision, Recall, và F1-score cho từng lớp (face, mask).

|  |  |
| --- | --- |
| **Thông số** | **Giá trị** |
| Epoch | 100 |
| Batch size | 32 |
| Learning rate | 0.01 - 0.001 |
| Early stopping | 20 |

*Bảng 3.2.1. Bảng tham số huấn luyện mô hình YOLO*

**U-NET**

Mô hình được huấn luyện với tập dữ liệu gồm khoảng 12.000 cặp ảnh mặt có và không khẩu trang, chia thành ba phần: 70% để huấn luyện, 20% để kiểm tra và 10% để kiểm định. Kích thước ảnh được chuẩn hóa về 256x256 px, chuẩn hóa về dải giá trị [0,1] để đảm bảo đầu vào nhất quán. Nhóm sử dụng hàm mất mát MAE cho mô hình này. Optimizer được sử dụng là Adam với learning rate ban đầu là 1e-4. Số epoch huấn luyện là 100 với batch size là 8. Ngoài ra, nhóm áp dụng kỹ thuật EarlyStopping để dừng sớm khi mô hình không còn cải thiện trên tập validation. Sau mỗi epoch, ảnh phục hồi từ tập kiểm tra sẽ được lưu lại để đánh giá chất lượng. Quá trình huấn luyện được theo dõi bằng biểu đồ PSNR và SSIM để đánh giá hiệu quả phục hồi.

|  |  |
| --- | --- |
| **Thông số** | **Giá trị** |
| Epoch | 100 |
| Batch size | 8 |
| Learning rate | 0.00001 |
| Early stopping | 10 |

*Bảng 3.2.2 Bảng tham số huấn luyện mô hình U-Net*

# **CHƯƠNG IV. ĐÁNH GIÁ KẾT QUẢ MÔ HÌNH**

# **4.1 Đánh giá kết quả mô hình YOLO**

## **4.1.1 Đánh giá loss của epoch cuối cùng**

- Box Loss = 0.3158: Đây là CIoU Loss, đo lường sai số trong việc dự đoán vị trí và kích thước của các hộp giới hạn (bounding box). Giá trị 0.3158 khá thấp, cho thấy mô hình dự đoán vị trí hộp giới hạn chính xác.

- Cls Loss = 0.1583: Đây là Binary Cross-Entropy Loss, đo lường sai số trong việc phân loại đối tượng (face hoặc mask). Giá trị 0.1583 rất thấp, cho thấy mô hình phân loại 2 lớp một cách hiệu quả.

- Dfl Loss = 0.9505: Distribution Focal Loss đo lường sai số trong việc dự đoán phân phối vị trí hộp giới hạn. Giá trị này cao hơn so với Box Loss và Cls Loss, nhưng vì Dfl Loss thường có thang giá trị lớn hơn và tập trung vào việc cải thiện độ chính xác của hộp.

## **4.1.2 Đánh giá trên tập Validation và Test**

Mô hình YOLOv8 được huấn luyện và đánh giá cho bài toán nhận diện gương mặt và khẩu trang đạt hiệu suất tốt trên cả tập validation và test. Các chỉ số đánh giá cho thấy độ chính xác cao với mAP@0.5 đạt 0.995 và mAP@0.5:0.95 khoảng 0.946 trên cả hai tập, thể hiện khả năng phát hiện và định vị chính xác. Precision và Recall của lớp face và mask đều gần 1.0 (Precision: 0.999–1.0, Recall: 0.998–1.0), với F1-score khoảng 0.999, cho thấy mô hình cân bằng tốt giữa việc giảm thiểu dự đoán sai và bỏ sót. Các giá trị loss cuối cùng (Box Loss, Cls Loss, Dfl Loss) khẳng định mô hình hội tụ tốt, dù Dfl Loss hơi cao, có thể cải thiện bằng cách bổ sung dữ liệu phức tạp.

|  |  |  |
| --- | --- | --- |
| **Độ đo** | **Validation** | **Test** |
| mAP@0.5 | 0.9950 | 0.9950 |
| mAP@0.5:0.95 | 0.9456 | 0.9464 |
| Precision lớp face | 1.0000 | 0.9992 |
| Recall lớp face | 0.9978 | 0.9986 |
| F1-score lớp face | 0.9989 | 0.9989 |
| Precision lớp mask | 0.9995 | 0.9991 |
| Recall lớp mask | 0.9983 | 1.0000 |
| F1-score lớp mask | 0.9989 | 0.9995 |

*Bảng 4.1.2. Giá trị các metrics trên tập Validation và Test*

A group of graphs showing the value of a number

AI-generated content may be incorrect.

Hình 4.1.2. Biểu đồ kết quả huấn luyện

**4.1.3 Hình ảnh dự đoán**

Kết quả đánh giá mô hình dựa trên các chỉ số của các metric rất tốt cho thấy mô hình có khả năng nhận diện khuôn mặt và khẩu trang rất chính xác. Mô hình sau khi train có dung lượng khá thấp (khoảng 6 MB) và tốc độ inference từ 10.5ms ~ 10.9ms phù hợp cho các hệ thống nhẹ.

A person wearing a face mask

AI-generated content may be incorrect.

Hình 4.1.3 Hình ảnh dự đoán của mô hình sau khi train

* 1. **Đánh giá kết quả mô hình U-Net**

### **4.2.1 Thang đo SSIM, PSNR, LPIPS và MAE**

* **SSIM (Structural Similarity Index):** 0.8209 ± 0.0631, cho thấy mô hình bảo tồn khá tốt cấu trúc và chi tiết của ảnh gốc trong quá trình tái tạo lại khuôn mặt không bị che.

*A graph of a graph

AI-generated content may be incorrect.*

Hình 4.2.1 Biểu đồ kết quả huấn luyện với SSIM và PSNR

* **PSNR (Peak Signal-to-Noise Ratio):** 23.49 ± 3.08, phản ánh chất lượng ảnh phục hồi ở mức độ ổn trong bài toán xử lý ảnh, khi giá trị PSNR càng cao thì ảnh càng ít nhiễu so với ảnh gốc.
* **LPIPS (Learned Perceptual Image Patch Similarity):** 0.2159 ± 0.0526, là thang đo khoảng cách cảm nhận giữa ảnh gốc và ảnh phục hồi. Giá trị LPIPS thấp cho thấy ảnh đầu ra gần giống với ảnh thật về mặt cảm nhận thị giác.
* **MAE (Mean Absolute Error):** 0.0360 ± 0.0148, xác nhận lại mức độ sai lệch trung bình thấp giữa ảnh gốc và ảnh được phục hồi.

A graph of a graph

AI-generated content may be incorrect.

Hình 4.2.1.2 Biểu đồ kết quả huấn luyện với Loss, MAE, MSE

### **4.2.2 Hình ảnh đầu ra**

Mặc dù các chỉ số đánh giá định lượng như SSIM, PSNR, LPIPS và MAE cho thấy mô hình đạt kết quả khá tốt về mặt tái tạo hình ảnh, tuy nhiên, khi quan sát trực tiếp các ảnh được phục hồi sau dự đoán, vẫn tồn tại một số hạn chế đáng chú ý. Cụ thể, các đặc điểm khuôn mặt quan trọng như mũi và miệng – vốn đóng vai trò then chốt trong việc nhận diện khuôn mặt – thường không được tái hiện rõ nét hoặc bị mờ nhòe, làm giảm tính chân thực và tự nhiên của ảnh đầu ra.

A person with blue face

AI-generated content may be incorrect.

Hình 4.2.2 Kết quả ảnh đầu ra của U-Net

Có thể thấy rằng mặc dù mô hình có khả năng khôi phục tổng thể khuôn mặt khá ổn định, nhưng vẫn chưa học được đầy đủ các đặc trưng cục bộ quan trọng về hình thái học khuôn mặt. Nguyên nhân có thể xuất phát từ kiến trúc mô hình chưa đủ tinh vi trong việc học đặc trưng chi tiết, hoặc bộ dữ liệu chưa cung cấp đủ đa dạng về vị trí và hình dạng khẩu trang để mô hình học hiệu quả hơn trong các vùng quan trọng như vùng miệng – mũi.

Vì vậy, trong các bước cải tiến tiếp theo, nhóm sẽ xem xét các hướng phát triển mô hình như: tập trung trọng số loss cao hơn ở khu vực bị che bởi khẩu trang nhằm giúp mô hình tập trung phục hồi các vùng quan trọng hoặc sử dụng mô hình chuyên biệt cho việc xây dựng mô hình nhận diện này tốt hơn.

# **CHƯƠNG V. TỔNG KẾT**

Sau khi đánh giá các thông số của hai mô hình, Hệ thống nhận diện khẩu trang và tái tạo khuôn mặt có thể được triển khai hiệu quả thông qua việc kết hợp hai mô hình học sâu chuyên biệt: YOLO cho tác vụ phát hiện khuôn mặt và khẩu trang, cùng với Unet cho nhiệm vụ tái tạo lại khuôn mặt bị che khuất. Trong quá trình huấn luyện và đánh giá, mô hình YOLO cho thấy hiệu quả vượt trội với các chỉ số đánh giá ấn tượng. Điều này chứng tỏ mô hình có khả năng nhận diện khuôn mặt và khẩu trang rất chính xác, ổn định và hoàn toàn có thể được ứng dụng trực tiếp trong các bài toán thực tế như giám sát an ninh, điểm danh không tiếp xúc hay kiểm tra tuân thủ quy định đeo khẩu trang.

Mặt khác, mô hình Unet – được sử dụng để dự đoán và tái tạo lại phần khuôn mặt bị che bởi khẩu trang – cũng đạt được các kết quả đánh giá khá tốt thông qua các chỉ số như PSNR và SSIM. Tuy nhiên, trong thực tế, các ảnh được tái tạo vẫn chưa hoàn toàn rõ nét và chính xác như mong muốn. Một số chi tiết trên khuôn mặt, đặc biệt là vùng bị che phủ bởi khẩu trang, vẫn còn mờ hoặc không tự nhiên, cho thấy mô hình hiện tại vẫn cần được cải thiện thêm. Những yếu tố như chất lượng dữ liệu huấn luyện, độ đa dạng khuôn mặt và kỹ thuật tái tạo ảnh có thể ảnh hưởng đến chất lượng đầu ra, và sẽ là những hướng phát triển quan trọng trong tương lai.

Tóm lại, việc kết hợp giữa mô hình phát hiện (YOLO) và mô hình tái tạo (Unet) là một hướng đi đầy tiềm năng cho các ứng dụng thị giác máy tính liên quan đến nhận diện và phục hồi khuôn mặt. Hệ thống bước đầu đã cho thấy hiệu quả và khả năng ứng dụng cao, tuy nhiên vẫn còn không gian để tối ưu nhằm đạt được kết quả tốt hơn trong các tình huống thực tế phức tạp hơn.

# **TÀI LIỆU THAM KHẢO**

[1] Hosen, M. D., & Baek, J. (2022*). Mask-to-Face: Realistic Face Image Generation from Masked Face Input.* arXiv preprint arXiv:2209.08850.

[2] STRV Engineering. (2021, March 2). *Mask2Face: How We Built AI That Shows Face Beneath Mask.* STRV Blog.

[3] Hosen, M. D. (2022). *Mask Face Inpainting Using Residual Attention UNet* [Source code]. GitHub. https://github.com/mdhosen/mask-face-inpainting-using-residual-attention-unet

[4] Hosen, M. D., & Baek, J. (2022). *HiMFR: A Hybrid Masked Face Recognition Through Face Inpainting*. ResearchGate.