

# Nhận diện người đeo khẩu trang và không đeo khẩu trang sử dụng Deep Learning

Trịnh Linh Chi<sup>1</sup>, Phạm Thảo Nhi<sup>2</sup>, Nguyễn Hữu Trường<sup>3</sup>

Đại Học Công Nghệ Thông Tin ĐHQG Tp.HCM, VN  
{19521285, 19520815, 18521564}@uit.edu.vn

**Tóm tắt nội dung** Hiện tại, Corona virus với nhiều loại biến thể khác nhau đã và đang dần phát triển, lây nhiễm và lan rộng, gây ảnh hưởng đến sức khỏe và tính mạng người dân trên nhiều quốc gia khắp các châu lục, trở thành mối nguy hiểm đáng lo ngại trên toàn thế giới. Việc thực hiện các chiến lược ngăn chặn lây lan virus để giảm thiểu tác động xấu được đặc biệt chú ý cao độ. Dưới tình hình hiện tại, vẫn chưa có thuốc kháng lại dịch bệnh hiệu quả, cùng với sự hạn chế về nguồn lực y tế ở các quốc gia nghèo, nhiều biện pháp đã được WHO đề xuất để kiểm soát tỉ lệ lây nhiễm và tránh cạn kiệt nguồn lực. Trong đó, đeo khẩu trang là biện pháp can thiệp không sử dụng thuốc men khả thi và tiết kiệm nhất. Vì vậy, mọi quốc gia đều bắt buộc người dân của mình đeo khẩu trang khi ở nơi công cộng. Để đóng góp cho công cuộc bảo vệ sức khỏe cộng đồng. Đồ án chúng em được xây dựng với mục đích giúp tạo ra kỹ thuật có thể phát hiện và cảnh báo những người không đeo khẩu trang ở nơi công cộng với độ chính xác cao và số lượng đối tượng nhiều đáng kể. Việc thực hiện được tiến hành với hai mô hình object detection có hiệu suất cao trong thời gian gần đây là Yolov4 và Yolov5. Các mô hình được huấn luyện trên dữ liệu thực tế được chúng em tự tổng hợp và gán nhãn. Mô hình thu được đã thể hiện kết quả khá tốt trên các video sử dụng để thử nghiệm.

**Keywords:** Nhận diện đeo khẩu trang · Ngăn chặn corona virus · Yolo

## 1 Giới thiệu

Theo báo cáo của bộ y tế [1], từ 16h ngày 22/01 đến ngày 23/01 trên hệ thống Quốc gia quản lý ca bệnh Covid-19 ghi nhận 14.978 ca nhiễm mới, trong đó có 44 ca nhập cảnh và 14.934 ca ghi nhận trong nước tại 62 tỉnh, thành phố. Trong đó, 10.324 ca nằm trong cộng đồng. Toàn quốc tổng cộng có 2.141.422 ca, với 1.804.849 ca qua khỏi và 36.719 ca tử vong.

Nhiều nghiên cứu cũng như chính Giám đốc tổ chức y tế Liên Mỹ (PAHO) [2] đã chỉ ra rằng việc đeo khẩu trang là cần thiết và có thể giảm thiểu được rất nhiều xác suất lây nhiễm, đồng thời khuyến khích cá nhân và tổ chức tuân thủ chặt chẽ quy định đeo khẩu trang nơi công cộng. Theo đó, việc cần thiết để phát triển, lắp đặt các hệ thống cảnh báo thực hiện việc đeo khẩu trang là không thể bàn cãi.

Việc nhận diện một người có đeo khẩu trang hay không, trên thực tế chính là bài toán phát triển ngược từ bài toán nhận diện khuôn mặt. Trong quá khứ, đã có nhiều thành quả của bài toán từ những mô hình học máy phức tạp với kết quả thấp [3] [4], hay những mô hình CNN với kết quả cải thiện rõ rệt [5], [6].

Đồ án của chúng em thừa hưởng kết quả của từ các công trình trước, sử dụng và tinh chỉnh các mô hình object detection hiệu suất cao trong thời gian gần đây trên bộ dữ liệu tự thu thập và gán nhãn thủ công. Những đóng góp bao gồm:

1. Thông tin bộ dữ liệu được trình bày ở phần 3.
2. Nền tảng lý thuyết các mô hình sử dụng được trình bày ở phần 4
3. Độ đo đánh giá và kết quả thực nghiệm được trình bày ở phần 5
4. Kết luận và hướng phát triển được trình bày ở phần 6

## 2 Các nghiên cứu liên quan

Học mẫu (Pattern learning) và nhận dạng đối tượng (Object recognition) là hai task cơ bản trong Computer Vision. Object recognition bao gồm hai task là phân loại ảnh (Image classification) và nhận dạng đối tượng (Object detection) [7]. Quy trình object recognition (pipeline) bao gồm việc tạo ra các vùng và phân lớp các vùng được tạo [8].

Các mô hình nhận diện single-stage coi việc phát hiện các region proposals như một bài toán hồi quy đơn giản, với OverFeat và DeepMultiBox là những ví dụ ban đầu. YOLO (You Only Look Once) đã phổ biến phương pháp tiếp cận single-stage khi đạt được tốc độ nhận diện đáng kể nhưng lại có độ chính xác thấp khi so sánh với các mô hình nhận diện two-stage; đặc biệt là đối với các đối tượng nhỏ [9].

Trái ngược với bộ nhận diện single-stage, bộ nhận diện two-stage dự đoán các đề xuất trong một hình ảnh và sau đó áp dụng bộ phân loại cho các khu vực này để phân loại phát hiện tiềm năng. Mạng nơ-ron tích chập dựa trên vùng còn được viết tắt là R-CNN [10] được mô tả vào năm 2014 bởi Ross Girshick và cộng sự. SPPNet thu thập các đặc điểm từ các region proposals khác nhau và đưa vào một lớp được kết nối đầy đủ để phân loại. Tiếp theo, Fast-R-CNN là một phần mở rộng trên R-CNN và SPPNet [?], [7]. Mặc dù Fast-R-CNN tích hợp hiệu quả các điểm mạnh của R-CNN và SPPNet nhưng vẫn chậm hơn so với các máy dò single-stage.

## 3 Dataset

Trong phần này, chúng em sẽ mô tả về bộ dữ liệu mà chúng em đã sử dụng trong bài báo cáo này.

Bộ dữ liệu bao gồm 3200 bức ảnh, trong đó:

- 1000 bức ảnh từ cuộc thi Data-Centric AI Competition FPT <sup>1</sup>.

---

<sup>1</sup> Data-Centric AI Competition FPT: <https://www.datacomp.io/trang-chu>

- 2200 bức ảnh được nhóm chúng em thu thập từ hình ảnh và video trên internet, sau đó được gán nhãn thủ công bằng tool Labelme<sup>2</sup>.

Nhãn của dữ liệu sau khi gán sẽ được chuyển từ file JSON về file Txt theo form yolo tương ứng thủ công. Với form yolo bao gồm nhãn, tọa độ điểm trung tâm, chiều dài, chiều rộng của bounding box. Các đối tượng sẽ được gán nhãn nếu có thể nhìn thấy trên 50% đối tượng.

Bộ dữ liệu được chia thành 3 tập dữ liệu: train (tập huấn luyện), val (tập xác thực), test (tập kiểm tra). Dữ liệu của các tập train, val, test gồm các ảnh được gán nhãn bounding box theo 2 lớp:

- Deo khẩu trang – được ký hiệu là 1.
- Không deo khẩu trang – được ký hiệu là 0.
- Đối với những đối tượng deo khẩu trang sai (tức là deo khẩu trang dưới mũi) thì xem như không deo khẩu trang, nhãn sẽ được gán là 0.
- Các đối tượng nhỏ và mờ (không nhìn rõ mặt mũi) sẽ không được gán nhãn.



**Hình 1.** Ví dụ về ảnh đã được gán nhãn bounding box trong tập train.

Trong đó, tập train gồm 2406 ảnh, tập val gồm 406 ảnh và tập test gồm 388 ảnh với số lượng mỗi nhãn trong các tập có số lượng như trong Bảng ??.

	0	1
Train	5172	6499
Val	309	1176
Test	903	1348

**Bảng 1.** Số lượng của mỗi nhãn trong các tập train, val, test.

---

<sup>2</sup> Link tool Labelme: <https://github.com/wkentaro/labelme>

## 4 Model

Trong bài báo cáo này, chúng em sẽ sử dụng hai mô hình là YOLOv4 và YOLOv5 để nhận diện người đeo khẩu trang và không đeo khẩu trang trong video.

### 4.1 YOLOv4

Thuật toán YOLOv4 về cơ bản cũng thừa kế các phương pháp cơ bản của các phiên bản trước của YOLO, tuy nhiên YOLOv4 áp dụng một số thuật toán phát hiện vật thể nhanh, tối ưu hóa các phép toán thực hiện song song giúp tăng tốc độ nhận diện và tăng độ chính xác.

Cấu trúc nhận diện vật thể của YOLOv4 thường có:

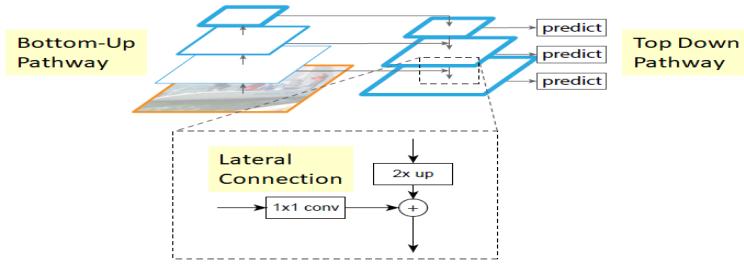
**Backbone:** là mô hình pre-trained của 1 mô hình transfer learning dùng để học các đặc trưng và vị trí của vật thể. Các mô hình transfer learning thường là VGG16, ResNet-50, ResNeXt-101, Darknet53,... Ở đây mô hình transfer learning được áp dụng trong YOLOv4 của chúng em là CSPDarknet53.

**Neck:** Neck thường được dùng để làm giàu thông tin bằng cách kết hợp thông tin giữa quá trình bottom-up và quá trình top-down (do có một số thông tin quá nhỏ khi đi qua quá trình bottom-up bị mất mát nên quá trình top-down không tái tạo lại được). Các mô hình được dùng trong quá trình Neck của YOLOv4 là SPP, PAN.

Trong quá trình backbone, khi bottom-up stream sẽ giúp việc xác định vị trí của vật thể nhanh và chính xác, tuy nhiên khi đi qua quá trình trên việc các feature map càng ngày càng thu nhỏ, độ phân giải giảm xuống làm cho một số các vật thể nhỏ qua đó cũng bị biến mất làm mất thông tin. Để khắc phục điều đó trước khi đẩy về head để nhận dạng, Neck thường thực hiện một quá trình ngược lại (top-down stream) tạo ra các reconstruction map để khôi phục lại một số thông tin bị mất trong quá trình bottom-up stream. Nhưng một số thông tin đã bị mất khi khôi phục bằng quá trình top-down cũng không còn hiển thị lại nữa, do đó mạng FPN (Feature Pyramid Network), sẽ tái tạo lại các thông tin bị mất nhờ các kết nối skip-connection giữa các feature map được lấy tích chập 1x1 và reconstruction map tạo ra các feature map mới giàu thông tin giúp việc phát hiện và phân loại vật thể đạt độ chính xác cao hơn.

Trong YOLOv4, có 2 cải tiến đối với mô hình FPN gồm: Đối với khối skip-connection giữa feature map và reconstruction map, tác giả sử dụng mô hình YOLO-SPP (Spatial Pyramid Pooling) thay cho các feature map được tích chập 1x1. YOLO-SPP được giới thiệu trong YOLOv3 và tiếp tục được sử dụng trong YOLOv4. YOLO-SPP sử dụng các mạng tích chập max-pooling với các kích thước filter khác nhau, sau khi được lấy tích chập các lớp ngõ ra sẽ được xếp chồng lên nhau. Việc thực hiện max-pooling như vậy giúp giữ được các đặc trưng quan trọng của feature map mà gần như không làm giảm tốc độ xử lý.

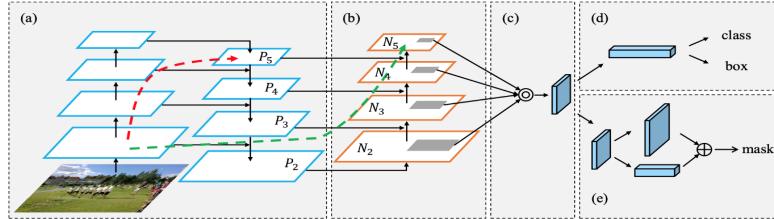
PAN là một cải tiến từ FPN. PAN thêm một đường bottom-up được gắn sau đường top-down của FPN. Ở đường bottom-up thứ hai này, feature map của



**Hình 2.** Mô hình FPN.

tầng trước đó sẽ được lấy tích chập qua filters  $3 \times 3$  rồi add với feature của tầng tương ứng ở đường top-down.

Tuy nhiên trong YOLOv4, thay vì add thì được thay bằng concat với tầng của đường top-down. Sau đó, các đối tượng sẽ được nhận diện ở từng tầng.



**Hình 3.** Mô hình mạng PAN.

**Head:** phần Head được sử dụng để tăng khả năng phân biệt đặc trưng để dự đoán class và bounding box. Ở phần Head có thể áp dụng 1 tầng hoặc 2 tầng: Tầng 1 là Dense Prediction: dự đoán trên toàn bộ hình với các mô hình RPN, YOLO, SSD,... Tầng 2 là Sparse Prediction dự đoán trên từng mảng được dự đoán có vật thể với các mô hình R-CNN series,...

#### 4.2 YOLOv5

Yolov5 ra đời chỉ vài ngày sau Yolov4, với 4 phiên bản khác nhau: 5s, 5m, 5l, 5x với độ chính xác tăng dần và tốc độ giảm dần. Trong đó 5x là phiên bản mà chúng em sử dụng. Cấu trúc của Yolov5 bao gồm các phần:

**Backbone:** được sử dụng để trích xuất đặc trưng quan trọng từ ảnh đầu vào. Giống với Yolov4, Yolov5 sử dụng CSP - Cross Stage Partial Networks làm backbone để trích xuất các đặc trưng giàu thông tin từ ảnh đầu vào. CSPNet giải quyết vấn đề về thông tin gradient lặp đi lặp lại trong Backbone với quy mô

lớn và tích hợp thay đổi gradient trong feature map, do đó làm giảm số lượng parameters và FLOPS của mô hình, không chỉ đảm bảo suy luận tốc độ và độ chính xác, nhưng cũng làm giảm kích thước mô hình. CSPNet đã thể hiện được sự cải thiện rõ rệt về thời gian thực thi đối với các mạng neural có độ sâu lớn, quy mô lớn.

**Neck:** để tạo feature pyramid giúp mô hình khái quát về độ mở rộng đối tượng tốt hơn. Feature pyramid sẽ giúp mô hình hoạt động tốt hơn trên những dữ liệu mới, dữ liệu lạ. Có nhiều dạng pyramid features được sử dụng với Yolov5 như FPN, BiFPN, PANet,... Ở đây, Yolov5 sử dụng PANet làm Neck model. PANet sử dụng cấu trúc feature pyramid mới là FPN với bottom-up path giúp cải thiện propagation của low-level features. Cùng lúc đó, tổng hợp feature pooling thích ứng liên kết feature grid và tất cả feature levels giúp rút được thông tin có ích từ mọi propagation từ mọi feature level khác nhau truyền thẳng đến subnetwork. PANet cải thiện việc sử dụng các local signal ở các layer thấp, giúp nâng cao độ chính xác trong việc detect vị trí của đối tượng.

**Head:** Dùng để thực hiện phần detection cuối cùng. Áp dụng anchor box lên feature để cho ra kết quả với output vectors bao gồm tên class dự đoán, xác suất và tọa độ bounding box. Head của yolov5 có tên là yolo layer với 3 kích thước khác nhau ( $18 \times 18$ ,  $36 \times 36$ ,  $72 \times 72$ ) giúp đạt được multi-scale prediction đối với các kích thước nhỏ, vừa và lớn của đối tượng. Tùy theo khoảng cách xa gần mà đối tượng, cụ thể là mặt người trong các camera giám sát có thể to hay nhỏ khác nhau. Multi-scale detection giúp mô hình có thể thích ứng và đưa ra dự đoán chính xác trong những trường hợp này.

## 5 Kết quả thực nghiệm

### 5.1 Metrics đánh giá object detection model

Mục tiêu của một Object detection model là Classification và Localization. Vì vậy phương pháp đánh giá một Object detection model phải kết hợp được hai mục tiêu này. Về bản chất mỗi class trong Object detection model có một giá trị AP riêng. Bằng cách lấy trung bình các AP, ta thu được mAP. Cách tính AP được thể hiện dưới đây:

#### A. Intersection over Union - IoU:

IoU là tỉ số giữa phần giao nhau và phần hợp nhau giữa Ground Truth Bounding Box (Bounding Box mà ta vẽ thủ công khi gán nhãn đối tượng) và Bounding Box mà model dự đoán ra được.

Giá trị IoU nằm trong khoảng [0; 1]. IoU = 0 nghĩa là model dự đoán sai hoàn toàn, IoU = 1 nghĩa là model dự đoán chính xác hoàn toàn. Thông thường, sẽ có một giá trị ngưỡng (threshold) của IoU để xác định Bounding Box dự đoán ra có được chấp nhận hay không? Threshold thường là 0.5 hoặc 0.75. Nếu  $\text{IoU} > \text{Threshold}$ , Bounding Box là hợp lệ và ngược lại.

#### B. Precision và Recall:

Precision và Recall được xác định thông qua công thức:

$$Precision = \frac{TP}{TP + FP}$$

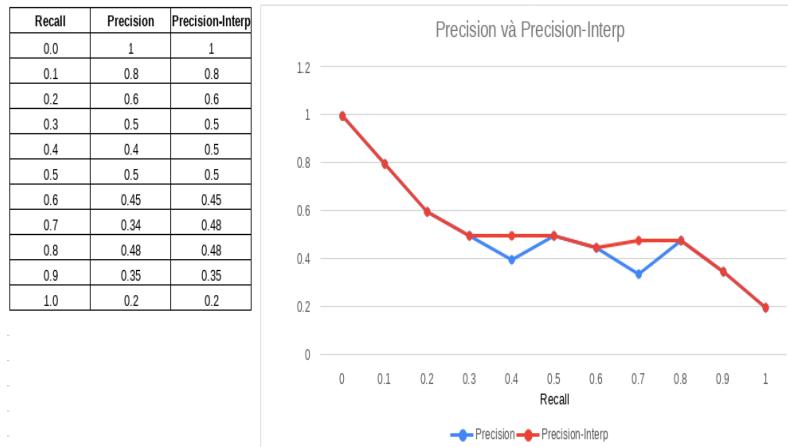
$$Recall = \frac{TP}{TP + FN}$$

Sử dụng IoU và Threshold, ta có thể xác định được TP, FP, TN, FN như sau:

- Nếu IoU  $\geq$  Threshold, Bounding Box được coi là TP - True Positive.
- Nếu IoU  $<$  Threshold, Bounding Box được coi là FP - False Positive.
- Nếu có đối tượng trong bức ảnh nhưng model không phát hiện được (không dự đoán được Bounding Box)  $\rightarrow$  FN - False Negative.
- Mọi phần trong bức ảnh mà không có đối tượng và model cũng dự đoán là không có đối tượng  $\rightarrow$  TN - True Negative.

### C. Tính toán mAP:

Biểu diễn recall theo precision ở dạng đồ thị theo ví dụ Hình 9. Ta sẽ suy ra được giá trị interpolation (nội suy) của precision được gọi là  $p_{interp}$ , là giá trị lớn nhất của precision nằm bên phải của nó. Tương ứng với recall từ 0 đến 1 (cách nhau 0.1) sẽ có 11 điểm recall. AP sẽ là trung bình  $p_{interp}$  của 11 điểm recall này.



**Hình 4.** Ví dụ về cách tính interpolation precision.

Cuối cùng mAP sẽ được tính bằng trung bình của tất cả AP của mỗi class:

$$mAP = \frac{1}{n} * \sum_{i=1}^n AP$$

## 5.2 Kết quả thực nghiệm

Với 408 hình ảnh từ tập test, kết quả thu được trên 2 model thông qua 3 độ đo precision, recall và mAP được thể hiện như trong Bảng 2 với confidence threshold là 0.4.

Theo độ đo mAP, trong cùng khoảng thời gian huấn luyện và cùng một bộ dữ liệu, mô hình yolov5 có kết quả 0.816, cao hơn so với mô hình yolov4 (0.73). Điều này có thể là vì cách training cũ theo darknet của yolov4 đòi hỏi nhiều thời gian huấn luyện hơn so với ultralytics v5 để cho ra kết quả tối ưu hơn.

Ngoài ra, nhóm chúng em còn tiến hành thử nghiệm mô hình detection trên một vài video mới và kết quả thu được khá khả quan cả về tốc độ, độ chính xác và khả năng multi-scaling detection.

Mô hình	Precision	Recall	mAP
YOLOv4	0.69	0.80	0.725542
YOLOv5	0.838	0.765	0.816

Bảng 2. Kết quả thực nghiệm

Sau đây là một số hình ảnh đầu ra của mô hình khi thử nghiệm trên tập test.

Hình 5, 6, 7, 8, 9 miêu tả chi tiết các trường hợp hình ảnh được mô hình nhận diện chính xác và đầy đủ các đối tượng đeo và không đeo khẩu trang.

Hình 10, 11 cho thấy các trường hợp mô hình nhận diện sai.



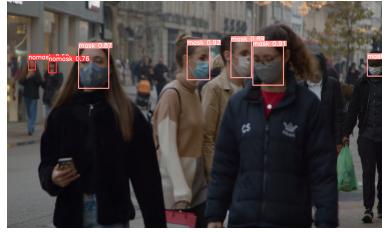
**Hình 5.** Bức ảnh đơn giản gồm hai đối tượng được mô hình dự đoán chính xác.



**Hình 6.** Bức ảnh gồm nhiều đối tượng gần nhau được nhận diện chính xác và đầy đủ.

## 6 Kết luận và hướng phát triển

Chúng em đã thành công thu thập, gán nhãn và xử lý được bộ dữ liệu thích hợp cho bài toán facemask detection với hơn 3000 ảnh. Chúng em đã cài đặt, huấn



**Hình 7.** Bức ảnh được mô hình nhận diện khá tốt dù các đối tượng đang đeo nhiều loại khẩu trang với các màu sắc và họa tiết khác nhau.



**Hình 8.** Bức ảnh gồm nhiều đối tượng với một phần khuôn mặt được dự đoán khá chính xác.



**Hình 9.** Mô hình cũng có thể nhận diện với các bức ảnh có độ sáng thấp với đối tượng nhận diện nhỏ.



**Hình 10.** Bức ảnh có đối tượng đeo khẩu trang sai bị nhận diện sai thành có đeo khẩu trang.



**Hình 11.** Bức ảnh với đối tượng chỉ lộ một phần nhỏ khuôn mặt bị nhận diện sai.

luyện và tối ưu mô hình facemask detection trên hai mô hình yolov4 và mô hình yolov5, thu được mô hình tốt nhất là yolov5 với kết quả 0.82 trên độ đo mAP. Đồng thời, thực hiện được prediction trên video với độ chính xác cao, tốc độ xử lý tốt và khả năng multi-scaling linh hoạt. Độ chính xác mô hình vẫn chưa thực sự tối ưu, cần phải cải thiện thêm. Ngoài ra, vẫn chưa tham khảo sử dụng được các mô hình object detection khác ngoài Yolo. Cùng lúc đó, vẫn chưa xử lý được trường hợp đeo khẩu trang không đúng quy định (hở mũi, kéo xuống cằm, etc.). Dự tính trong tương lai nhằm nâng cao, cải thiện hiệu suất mô hình chúng em sẽ thử nghiệm và tinh chỉnh các tham số nhiều lần, kết hợp với việc augmentation cũng như thêm nhãn riêng cho các đối tượng đeo khẩu trang không đúng quy định. Bên cạnh đó, chúng em sẽ tham khảo và cài đặt thêm các mô hình object detection khác ngoài yolo để so sánh cũng như tham chiếu cho việc cải thiện. Cuối cùng, chúng em hy vọng có thể áp dụng được mô hình trên thiết bị giám sát real-time trong thực tế với tốc độ, độ chính xác cao, có thể hỗ trợ được công tác phòng chống dịch bệnh trong tương lai tại địa phương.

## Tài liệu

1. Cổng thông tin của Bộ y tế về đại dịch covid 19, <https://covid19.gov.vn/>.
2. Social distancing, surveillance, and stronger health systems as keys to controlling COVID-19 Pandemic, PAHO Director says - PAHO/WHO | Pan American Health Organization. (n.d.). <https://www.paho.org/en/news/2-6-2020-social-distancing-surveillance-and-stronger-health-systems-keys-controlling-covid-19>.
3. Nanni L., Ghidoni S., Brahnam S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recogn.* 2017;71:158–172 <https://doi.org/10.1016/j.patcog.2017.05.025>.
4. Y. Jia et al., Caffe: Convolutional architecture for fast feature embedding, in: MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia, 2014 <https://doi.org/10.1145/2647868.2654889>.
5. P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks, 2014.
6. Erhan D., Szegedy C., Toshev A., Anguelov D. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014. Scalable Object Detection using Deep Neural Networks; pp. 2147–2154.
7. Inamdar M., Mehendale N. Real-Time Face Mask Identification Using Facemasknet Deep Learning Network. *SSRN Electron. J.* 2020 <https://doi.org/10.2139/ssrn.3663305>.
8. Qiao S., Liu C., Shen W., Yuille A. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018. Few-Shot Image Recognition by Predicting Parameters from Activations.
9. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016, vol. 2016-Decem, pp. 779–788, doi: 10.1109/CVPR.2016.91.
10. R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based Convolutional Networks for Accurate Object Detection and Segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 38 (1) (2015) 142–158, <https://doi.org/10.1109/TPAMI.2015.2437384>.