**TRƯỜNG ĐẠI HỌC XÂY DỰNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**----🙣🕮🙡----**

****

**THỊ GIÁC MÁY TÍNH**

*Đề tài:*

*PHÁT HIỆN TƯ THẾ CON NGƯỜI*

Các thành viên

Phạm Hữu Hải 64265

Đào Minh Đức 1510365

Nguyễn Thanh Phương 1539665

Phí Hoàng Long - 1532064

***Giáo viên hướng dẫn :GV Thái Thị Nguyệt***

Contents

[I.Mô tả bài toán và các phương pháp 3](#_Toc138797962)

[Xác định tư thế 2D và 3D 3](#_Toc138797963)

[Phương pháp truyền thống (classical) 3](#_Toc138797964)

[Phương pháp dựa trên deep learning 4](#_Toc138797965)

[II.Xác định cơ thể người bằng deep learning trong OpenCV 4](#_Toc138797966)

[OpenPose 4](#_Toc138797967)

[Mô hình 4](#_Toc138797968)

[Bản đồ tin cậy và Part Affinity Fields 5](#_Toc138797969)

[Loss functions: 6](#_Toc138797970)

[Bản đồ tin cậy: 7](#_Toc138797971)

[Part Affinity Fields: 7](#_Toc138797972)

[Lưu ý 9](#_Toc138797973)

[MoveNet 9](#_Toc138797974)

[Độ chính xác mô hình 11](#_Toc138797975)

[Hàm mất mát của mô hình 14](#_Toc138797976)

[So Sánh về độ phức tạp giữa 2 mô hình 16](#_Toc138797977)

[1. Kiến trúc mô hình: 17](#_Toc138797978)

[2. Độ chính xác: 17](#_Toc138797979)

[3. Tính linh hoạt và tốc độ: 17](#_Toc138797980)

[III.Tài liệu tham khảo 18](#_Toc138797981)

# I.Mô tả bài toán và các phương pháp

Xác định tư thế con người (HPE) là một cách xác định và phân loại các khớp trong cơ thể con người.

Về cơ bản, đó là một cách để nắm bắt một tập hợp tọa độ cho từng khớp (cánh tay, đầu, thân, v.v.), được gọi là điểm chính có thể mô tả tư thế của một người. Kết nối giữa các điểm này được gọi là một cặp .

Kết nối được hình thành giữa các điểm phải có ý nghĩa, điều đó có nghĩa là không phải tất cả các điểm đều có thể tạo thành một cặp

Có ba cách tiếp cận để mô hình hóa cơ thể con người:

1. Mô hình dựa trên bộ xương
2. Mô hình dựa trên đường viền
3. mô hình dựa trên khối lượng

## Xác định tư thế 2D và 3D

Xác định tư thế có thể được thực hiện ở dạng 2D hoặc 3D. Trong xác định tư thế 2D, mục tiêu là dự đoán vị trí của các khớp trong mặt phẳng hình ảnh. Trong xác định tư thế 3D, mục tiêu là dự đoán vị trí của các khớp trong không gian ba chiều.

Các phương pháp xác định tư thế người có thể được chia thành hai loại chính: phương pháp truyền thống (classical) và phương pháp dựa trên deep learning.

### Phương pháp truyền thống (classical)

Phương pháp truyền thống dựa trên các mô hình toán học cấu hình để mô hình hóa định dạng và cấu trúc của cơ thể. Các phương pháp này thường dựa trên các giả định về dạng cơ thể và cấu trúc khớp, đồng thời sử dụng các thuật toán tối ưu để tìm kiếm vị trí của các khớp trong hình ảnh. Một số phương pháp truyền thống biến phổ biến bao gồm RANSAC, Mô hình đồ họa, Tối ưu hóa bầy đàn hạt và Monte Carlo tuần tự. Các phương pháp truyền thống này thường được áp dụng trong các bài toán đơn giản và có điều kiện ánh sáng và góc nhìn tốt.

Khung cấu trúc hình ảnh (PSF) thường được coi là một trong những phương pháp truyền thống để xác định tư thế của con người. PSF chứa hai thành phần:

1. Discriminator: Nó mô hình hóa khả năng một bộ phận nhất định có mặt tại một địa điểm cụ thể. Nói cách khác, nó xác định các bộ phận cơ thể.
2. Prior : Nó được gọi là lập mô hình phân phối xác suất theo tư thế bằng cách sử dụng đầu ra từ bộ phân biệt đối xử; tư thế được mô hình hóa phải thực tế.

Phương pháp dựa trên deep learning**.**

Phương pháp dựa trên deep learning sử dụng các mô hình CNN để học các đặc trưng của hình ảnh và xác định vị trí của các khớp cơ thể. Các mô hình deep learning được huấn luyện trên các tập dữ liệu lớn, nơi nhãn đánh dấu là vị trí của các khớp cơ thể trong hình ảnh hoặc video. Các phương pháp deep learning có khả năng tổng quát hóa tốt hơn đối với các tư thế và hình dạng khác nhau của cơ thể. Một số phương pháp deep learning phổ biến bao gồm Convolutional Pose Machines (CPMs), Hourglass Networks, và OpenPose.

Để giải quyết những vấn đề này, các nhà nghiên cứu đã đưa ra hai cách tiếp cận:

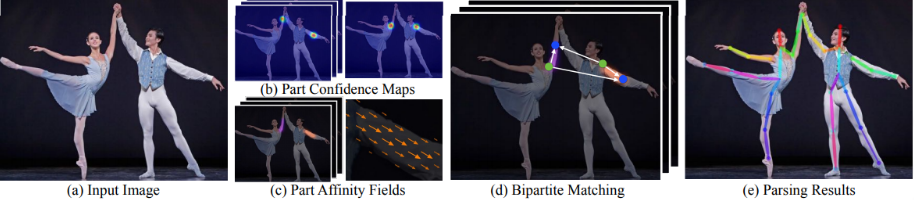
1. Từ trên xuống : Định vị con người trong hình ảnh hoặc video, sau đó xác định các phần, sau đó tính toán tư thế.
2. Từ dưới lên : Xác định các bộ phận cơ thể người trong ảnh, sau đó tính toán t

# II.Xác định cơ thể người bằng deep learning trong OpenCV

## OpenPose

OpenPose là hệ thống nhiều người theo thời gian thực đầu tiên cùng phát hiện các điểm chính trên cơ thể, tay, mặt và chân của con người (trong tổng số 135 điểm chính) trên các hình ảnh đơn lẻ. Nó được đề xuất bởi các nhà nghiên cứu tại Đại học Carnegie Mellon. Họ đã phát hành dưới dạng mã Python và Plugin Unity. Các tài nguyên này có thể được tải xuống từ kho lưu trữ [OpenPose](https://github.com/CMU-Perceptual-Computing-Lab/openpose)

### Mô hình

****

* Ở bước đầu tiên, hình ảnh được truyền qua mạng CNN cơ sở để trích xuất các bản đồ đặc trưng của đầu vào Trong bài báo. Trong bài báo này, các tác giả đã sử dụng 10 lớp đầu tiên của mạng VGG-19.
* Sau đó, bản đồ tính năng được xử lý trong một quy trình CNN nhiều giai đoạn để tạo Bản đồ tin cậy một phần và Trường mối quan hệ một phần
  + Bản đồ tin cậy một phần:
  + Một phần Part Affinity Fields
* Ở bước cuối cùng, Bản đồ tin cậy và Trường sở thích bộ phận  được tạo ở trên được xử lý bằng thuật toán so khớp lưỡng cực tham lam để có được tư thế cho từng người trong ảnh.

### Bản đồ tin cậy và Part Affinity Fields

* **Bản đồ tin cậy:** Bản đồ tin cậy là biểu diễn 2D của niềm tin rằng một bộ phận cơ thể cụ thể có thể được định vị trong bất kỳ pixel cụ thể nào. Bản đồ tin cậy được mô tả bằng phương trình sau:

https://lh3.googleusercontent.com/QEL9hHMbG5-gpDgcEJWpM-domFnJYazc5py5YtUxZhG8GctSCPXEjyQ6Wuy0KPgA-RNxp9FgL0OpdeLUwBUsrDHz1q7aKWxsyInqpZ8pLc4nwtm_RJdfe1UfBK2rPgc6rJ_U8Yr9t8fjsdVqSTPEfn0

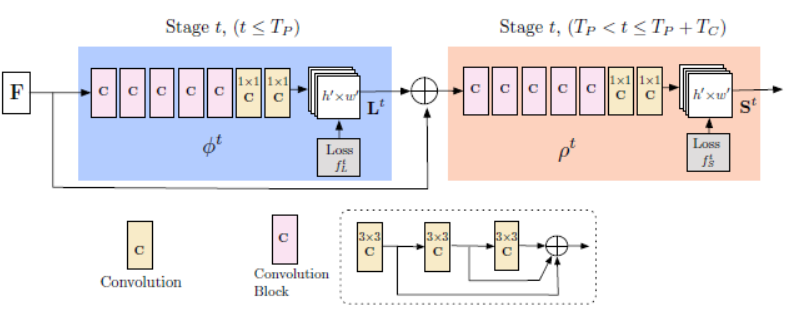
trong đó *J* là số vị trí bộ phận cơ thể.

**Part Affinity Fields:** Part Affinity là một tập hợp các trường vectơ 2D mã hóa vị trí và hướng của các chi của những người khác nhau trong ảnh. Nó mã hóa dữ liệu dưới dạng kết nối theo cặp giữa các bộ phận cơ thể.

https://lh6.googleusercontent.com/PZtVX5Msi63IClZ8648_0Zz787rAu4pExTFt7Yc_1d5i8AWzDUMc2mJjmTMcm_8rnUil6ecRY4Dvj_abHGVasNHkNOCwKf2wKSWvzFRTtPJPDBjpxoED_fVRHnXNftbSIxVZpJG36IS7dLLu30-ZGU8

**CNN nhiều giai đoạn:**

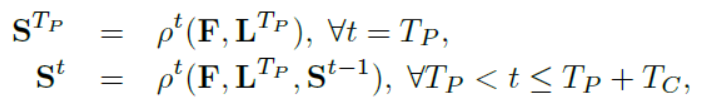
Kiến trúc đa CNN ở trên có ba bước chính:



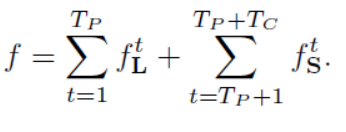
* Tập hợp các giai đoạn đầu tiên dự đoán Part Affinity Fields một phần tinh chỉnh L t  từ các bản đồ đặc trưng của mạng cơ sở F.

https://lh6.googleusercontent.com/fuR-NlxxJXCAbwad2LTV6L39w7pFTKLhhMo-7N2HyhcYH064zSn0422NOopASk7Na5DYfgyF4mTyK2N8sn_vcXaNAHsxyqYKhc4RZ7Gbs_XMf1rKtuxzw09mva-M7y1zezy_q4yBYfDlPrgIEDMmctE

* Tập hợp các giai đoạn thứ hai sẽ sử dụng Trường sở thích phần đầu ra từ các lớp trước đó để tinh chỉnh dự đoán phát hiện bản đồ tin cậy.

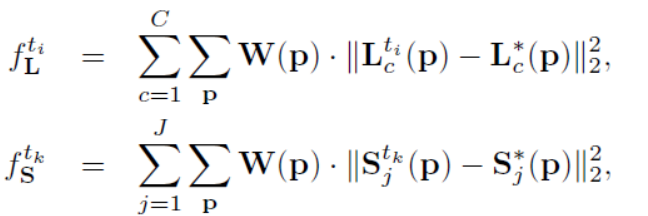


* Sau đó, S (bản đồ tin cậy) và L (Part Affinity Fields một phần) cuối cùng được chuyển vào thuật toán tham lam cho quá trình tiếp theo.



### Loss functions:

Hàm  *L2-loss* functionđược sử dụng để tính toán tổn thất giữa các bản đồ độ tin cậy được dự đoán và các trường Phần có mối quan hệ với các bản đồ và trường sự thật cơ bản

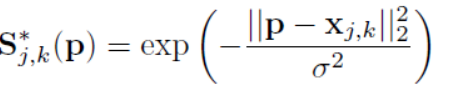


trong đó *L c \** là các Part Affinity Fields của phần chân lý nền, *S j \** là bản đồ độ tin cậy của phần chân lý nền và *W* là mặt nạ nhị phân với *W* ( *p* ) = 0 khi thiếu chú thích tại pixel p. Điều này là để ngăn chặn tổn thất thêm có thể được tạo ra bởi các mặt nạ này.

Giám sát trung gian ở mỗi giai đoạn được sử dụng để giải quyết vấn đề biến mất độ dốc bằng cách bổ sung độ dốc định kỳ.

### Bản đồ tin cậy:

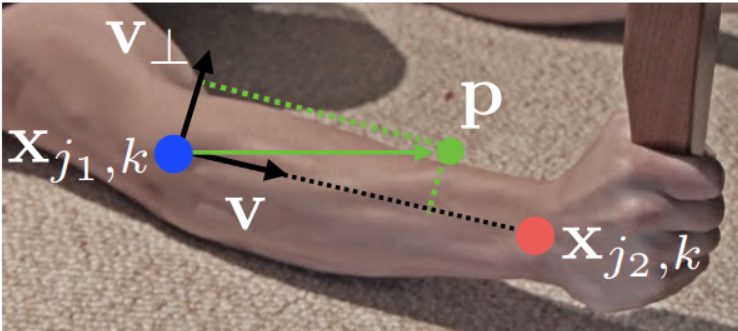
Bản đồ niềm tin cho mỗi người k và mỗi bộ phận cơ thể j được xác định bởi:



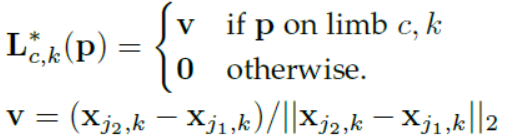
Đó là một đường cong Gaussian với những thay đổi dần dần trong đó *sigma* kiểm soát sự lan rộng của đỉnh. Đỉnh được dự đoán của mạng là tổng hợp các bản đồ độ tin cậy riêng lẻ của một toán tử tối đa.

### Part Affinity Fields:

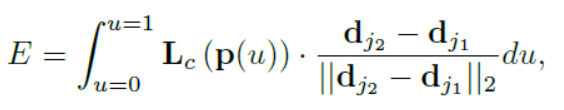
Part Affinity Fieldslà bắt buộc, đặc biệt là trong phát hiện tư thế nhiều người, chúng tôi bắt buộc phải ánh xạ các bộ phận cơ thể chính xác với cơ thể của nó. Bởi vì đối với nhiều người thì có nhiều đầu, nhiều tay, nhiều vai, v.v. Vì vậy, đôi khi chúng trở nên khó phân biệt khi chúng được nhóm chặt chẽ với nhau. PAF cung cấp kết nối giữa các phần khác nhau của cơ thể thuộc về cùng một người. Liên kết PAF mạnh hơn giữa các bộ phận cơ thể thể hiện khả năng cao các bộ phận cơ thể đó thuộc về cùng một người.



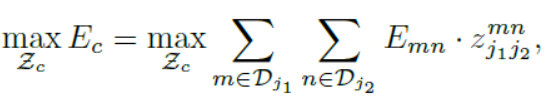
* Nếu p nằm trên nhánh thì *L\** là vectơ đơn vị nếu không thì nó bằng 0.



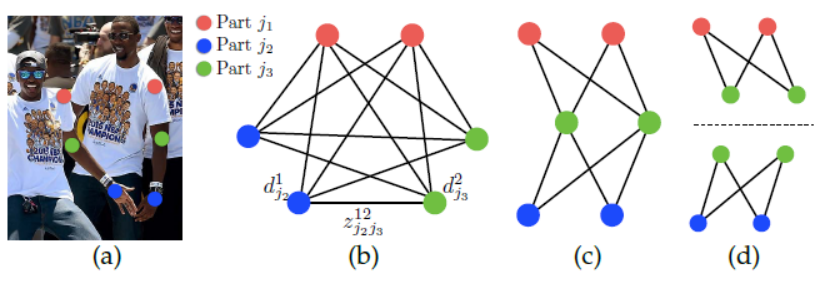
* Trường mối quan hệ Phần dự đoán, L c dọc theo đoạn thẳng là để đo độ tin cậy cho hai vị trí phần ứng cử dj 1 và dj 2 :



* Đối với nhiều người, Tổng E cần được tối đa hóa:

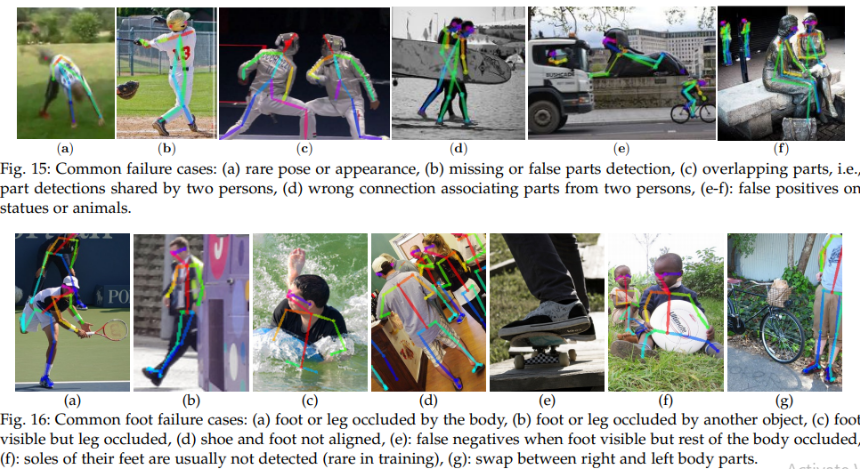


* Có nhiều cách tiếp cận để kết nối phần cơ thể như trong hình bên dưới:



* Liên kết bằng cách phát hiện bộ phận cơ thể.
* Liên kết bằng cách xem xét tất cả các cạnh và tạo biểu đồ k-partite
* Liên kếtbằng cách tạo cấu trúc cây.
* Liên kết bằng cách tạo biểu đồ hai bên khác nhau bằng thuật toán tham lam.

### Lưu ý

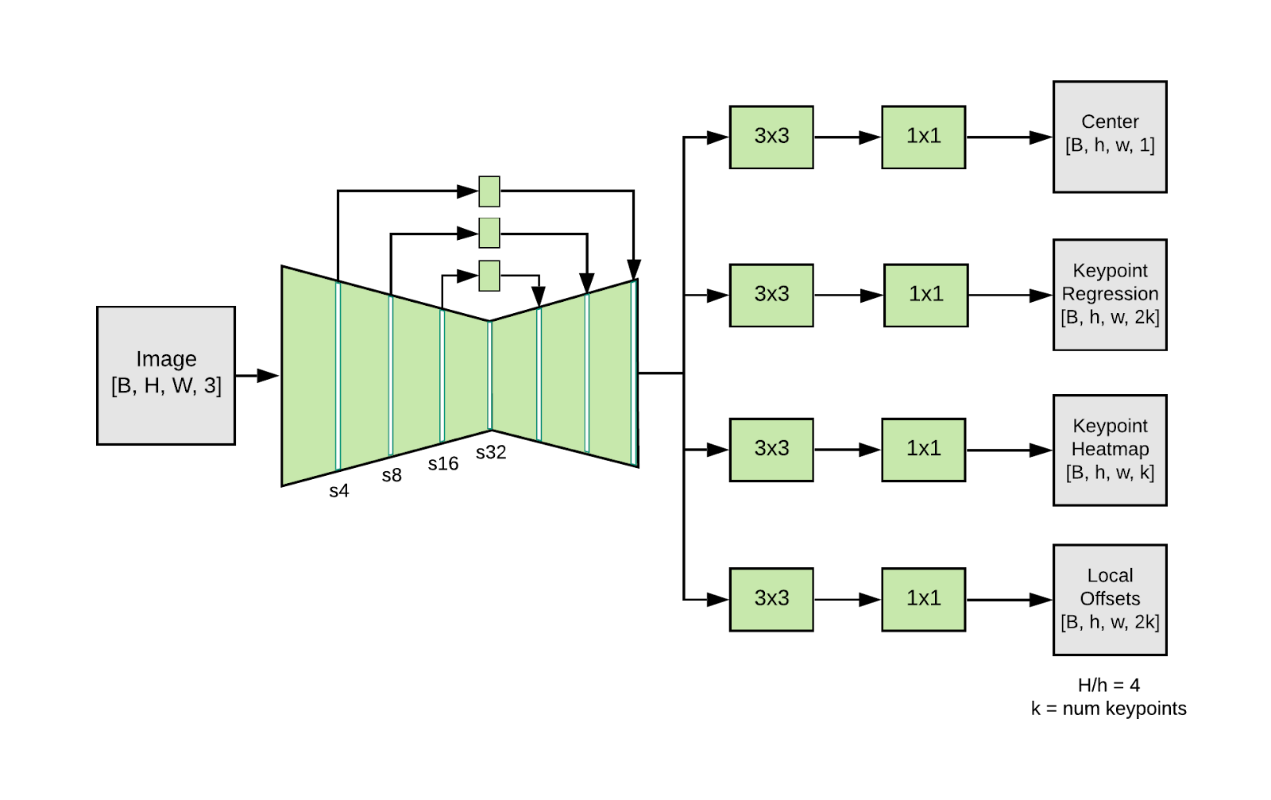
****

* OpenPose gặp sự cố khi ước tính tư thế khi ví dụ về sự thật trên mặt đất có các tư thế không điển hình và các ví dụ lộn ngược.
* Trong những hình ảnh có mật độ người tham gia cao, nơi mọi người chồng chéo lên nhau, cách tiếp cận có xu hướng hợp nhất các chú thích từ những người khác nhau, trong khi bỏ sót những người khác, do các PAF chồng chéo khiến việc phân tích cú pháp nhiều người tham lam không thành công

## MoveNet

MoveNet sử dụng bản đồ nhiệt để khoanh vùng các điểm chính của con người, hay còn gọi là mô hình ước tính từ dưới lên. Kiến trúc bao gồm một [bộ trích xuất đặc trưng](https://analyticsindiamag.com/how-feature-extraction-can-be-improved-with-denoising/) và một bộ [các đầu dự đoán](https://analyticsindiamag.com/pytorch-code-for-self-attention-computer-vision/) . Sơ đồ dự đoán cũng tuân theo [CenterNet](https://analyticsindiamag.com/object-detection-using-tensorflow/) , với những thay đổi đáng chú ý giúp cải thiện cả tốc độ và độ chính xác.

Trình trích xuất đặc trưng được sử dụng trong kiến ​​trúc MoveNet là MobileNetV2 với mạng kim tự tháp [feature pyramid network](https://arxiv.org/abs/1612.03144) ( [FPN](https://analyticsindiamag.com/detectron2/) ), cho phép xuất/kết quả bản đồ đặc trưng có độ phân giải cao, giàu ngữ nghĩa. Đối với đầu dự đoán, có bốn trong số chúng được gắn vào trình trích xuất đặc trưng, ​​dự đoán các trường hợp như:

* Bản đồ nhiệt trọng tâm người
* Trường hồi quy điểm chính
* Bản đồ nhiệt điểm chính của người
* Trường độ lệch 2D cho mỗi điểm khóa

Mặc dù những đầu ra này được tính toán song song, nhưng ta có thể hiểu rõ hơn về hoạt động của mô hình bằng cách xem xét chuỗi hoạt động sau:

Bước 1: Bản đồ nhiệt trung tâm người được sử dụng để xác định trung tâm của tất cả các cá nhân trong khung, được định nghĩa là trung bình cộng của tất cả các điểm chính thuộc về một người. Vị trí có điểm số cao nhất (được tính trọng số theo khoảng cách nghịch đảo từ trung tâm khung) được chọn.

Bước 2: Tập hợp các điểm chính ban đầu cho người được tạo bằng cách cắt đầu ra hồi quy điểm chính từ pixel tương ứng với trung tâm đối tượng. Vì đây là một dự đoán trung tâm – phải hoạt động trên các quy mô khác nhau – nên chất lượng của các điểm chính hồi quy sẽ không chính xác lắm.

Bước 3: Mỗi pixel trong bản đồ nhiệt của điểm chính được nhân với trọng số tỷ lệ nghịch với khoảng cách từ điểm chính hồi quy tương ứng. Điều này đảm bảo rằng chúng tôi không chấp nhận các điểm chính từ những người trong nền, vì chúng thường sẽ không ở gần các điểm chính bị hồi quy và do đó sẽ có điểm số thấp.

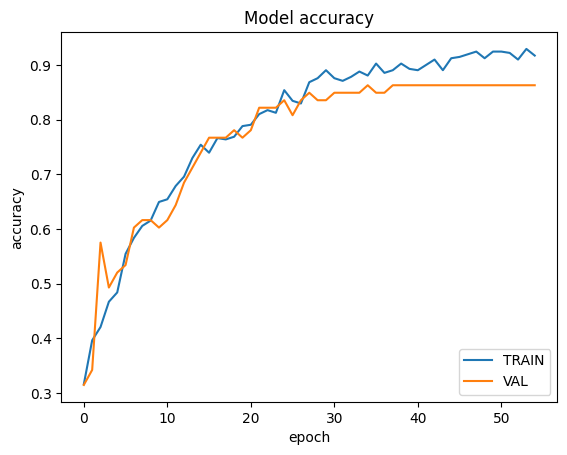
Bước 4: Tập hợp dự đoán điểm chính cuối cùng được chọn bằng cách truy xuất tọa độ của các giá trị bản đồ nhiệt tối đa trong mỗi kênh điểm chính. Sau đó, các dự đoán độ lệch 2D cục bộ được thêm vào các tọa độ này để đưa ra các ước tính tinh chỉnh. Xem hình bên dưới minh họa bốn bước này.

MoveNet đã được đào tạo về [COCO](https://cocodataset.org/#home) và [bộ dữ liệu nội bộ của Google](https://ai.googleblog.com/2018/09/building-google-dataset-search-and.html) có tên là Active. Mặc dù COCO chủ yếu phù hợp với các ứng dụng thể dục và khiêu vũ, thể hiện các tư thế thách thức và chuyển động mờ đáng kể, nhưng bộ dữ liệu Active được tạo ra bằng cách gắn nhãn các điểm chính (áp dụng 17 điểm chính cơ thể tiêu chuẩn của COCO) trên các video yoga, thể dục và khiêu vũ từ YouTube. Các chi tiết mô hình được tìm thấy [ở đây](https://storage.googleapis.com/movenet/MoveNet.SinglePose%20Model%20Card.pdf) .

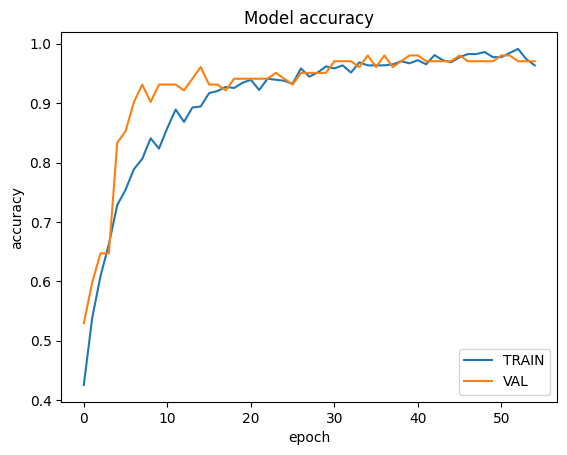
Từ khi được giới thiệu, MoveNet đã đạt được kết quả rất ấn tượng trong việc nhận diện hành động, đồng thời có khả năng chạy trên các thiết bị có tài nguyên hạn chế như điện thoại di động và máy tính nhúng. Điều này làm cho MoveNet trở thành một công nghệ tiềm năng trong việc tích hợp nhận diện hành động vào các ứng dụng di động và nhúng.

## Độ chính xác mô hình

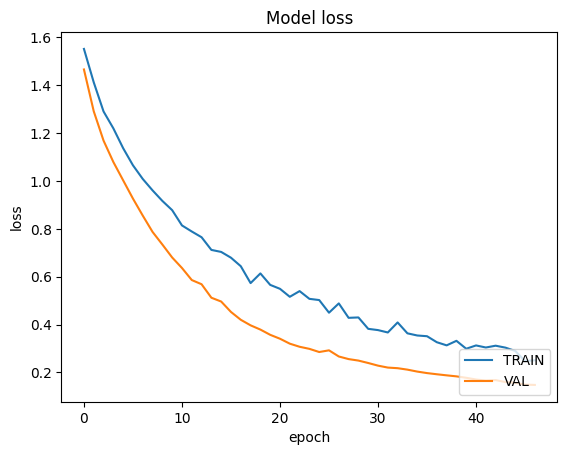
**OpenPose**

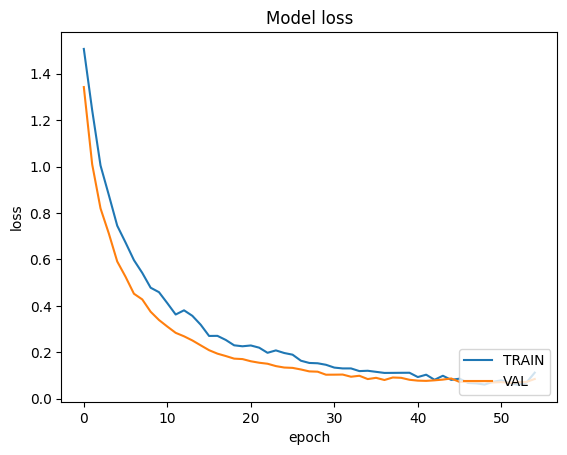
****

**MoveNet**

****

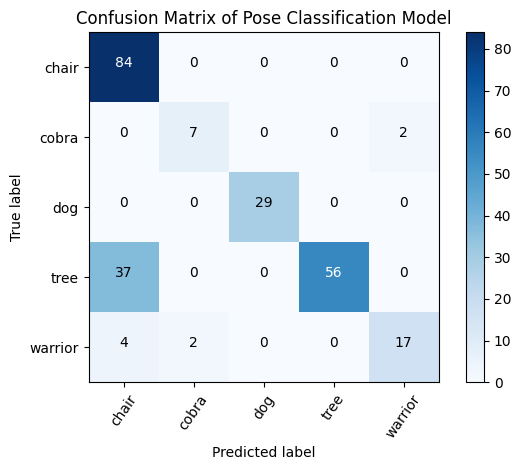
### Hàm mất mát của mô hình

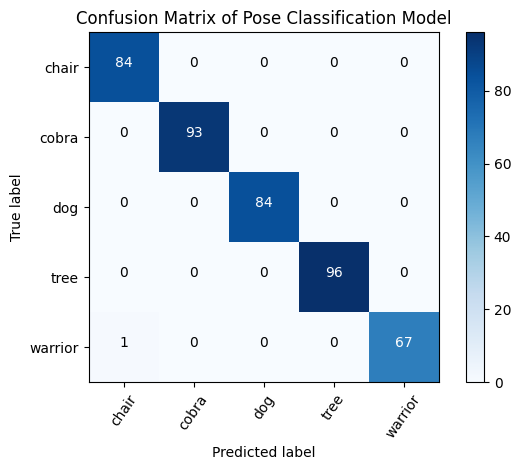
* **OpenPose**

* **MoveNet**

**Confution matrix**

**-OpenPose**

****

* **MoveNet**
* ****

## So Sánh về độ phức tạp giữa 2 mô hình

Độ phức tạp giữa Movenet và OpenPose có thể được so sánh theo một số yếu tố sau đây:

### 1. Kiến trúc mô hình:

   - Movenet: Movenet sử dụng một mạng nơ-ron tích chập (CNN) đơn giản và nhẹ để nhận diện và theo dõi các điểm khung hình trên cơ thể người. Mô hình Movenet thường có số lượng lớp và tham số nhỏ hơn so với OpenPose.

   - OpenPose: OpenPose sử dụng mạng nơ-ron tích chập (CNN) phức tạp hơn để trích xuất đặc trưng từ ảnh và sau đó sử dụng mô hình Part Affinity Fields (PAFs) để kết hợp các phần cơ thể thành các tư thế. Vì vậy, kiến trúc của OpenPose thường có số lượng lớp và tham số lớn hơn so với Movenet.

## 2. Độ chính xác:

   - Movenet: Do Movenet sử dụng một kiến trúc nhẹ hơn và đơn giản hơn, độ chính xác của nó có cao hơn so với OpenPose trong một số trường hợp phức tạp và yêu cầu khắt khe.

   - OpenPose: Với mạng nơ-ron tích chập (CNN) phức tạp và mô hình Part Affinity Fields (PAFs) để kết hợp thông tin, OpenPose có khả năng đạt được độ chính xác thấp hơn trong việc phát hiện và theo dõi tư thế con người.

## 3. Tính linh hoạt và tốc độ:

   - Movenet: Với mô hình nhẹ và tối ưu, Movenet có thể chạy trên các thiết bị có tài nguyên hạn chế như điện thoại di động và máy tính nhúng với tốc độ cao. Điều này giúp Movenet linh hoạt hơn trong việc triển khai trong các ứng dụng thời gian thực.

   - OpenPose: Do có một kiến trúc phức tạp hơn và số lượng lớp và tham số lớn hơn, OpenPose thường yêu cầu tài nguyên tính toán cao hơn và có thể chạy chậm hơn so với Movenet.

Tuy nhiên, độ phức tạp thực tế của Movenet và OpenPose cũng phụ thuộc vào cấu hình cụ thể của mỗi mô hình, bao gồm kích thước ảnh đầu vào, số lượng điểm khung hình (landmarks), và các yêu cầu cụ thể của ứng dụng.

# III.Tài liệu tham khảo

Dưới đây là một số tài liệu tham khảo về phát hiện tư thế con người sử dụng OpenPose:

1. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" - Các tác giả: Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh - Năm 2019. Đây là bài báo gốc giới thiệu mô hình OpenPose, phương pháp nhận dạng tư thế con người đa người thời gian thực sử dụng mạng nơ-ron tích chập (CNN) và mô hình Part Affinity Fields (PAFs).

2. "OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" - Các tác giả: Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh - Năm 2019. Đây là phiên bản công bố chính thức của OpenPose, bao gồm cả mã nguồn và mô hình huấn luyện.

3. "OpenPose: A Real-Time Multi-Person Keypoint Detection And Multi" - Tài liệu hướng dẫn sử dụng OpenPose được cung cấp trên trang chủ của dự án OpenPose (https://github.com/CMU-Perceptual-Computing-Lab/openpose). Tài liệu này cung cấp hướng dẫn chi tiết về cách sử dụng OpenPose để phát hiện tư thế con người trong ảnh và video.

4. "OpenPose: A Real-Time Multi-Person Key-point Detection And Tracking System For Body, Face, And Hands" - Bài báo công bố tại IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018. Bài báo này trình bày về phiên bản trước đó của OpenPose và đưa ra các cải tiến và kỹ thuật mới.

*-------------------------End-------------------------*