

# POSE ESTIMATION

Presented by: Thanh Duy



# 1.Pose Estimation

## 1.1 Định nghĩa

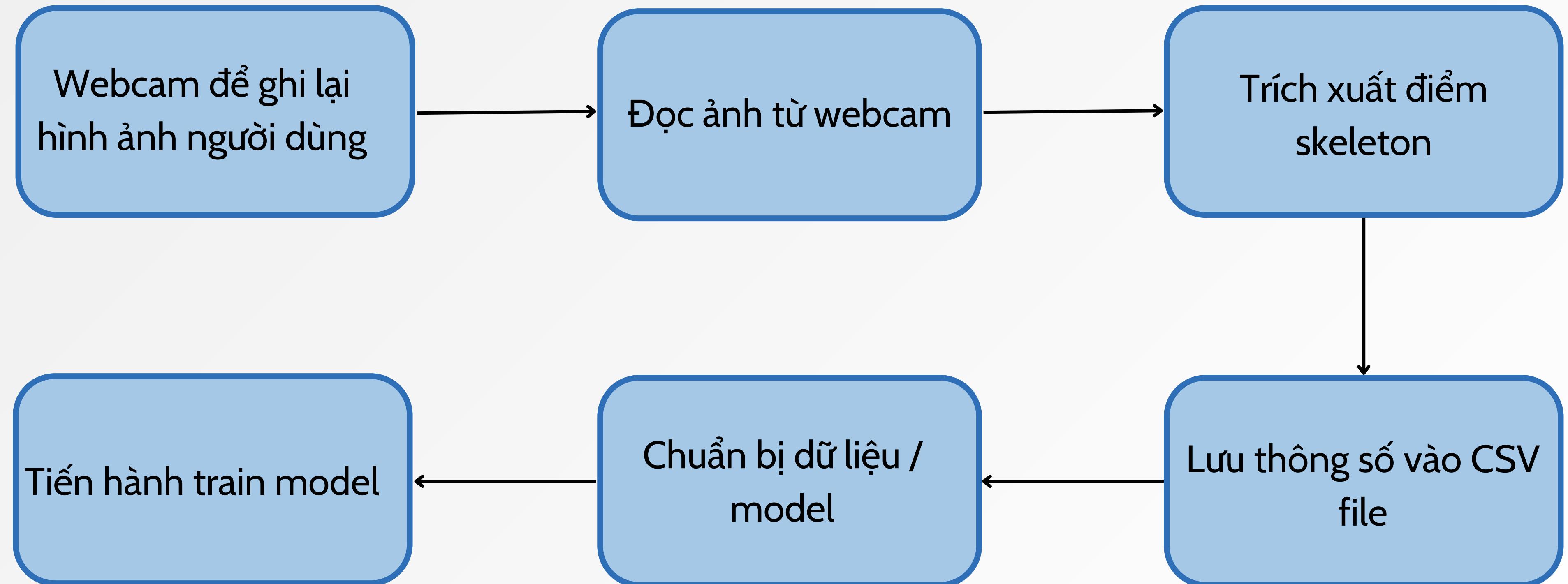
- Pose Estimation là một bài toán trong thị giác máy tính liên quan đến việc xác định cấu hình không gian của các bộ phận cơ thể người từ hình ảnh hoặc video.
- Công nghệ này nhằm mục đích xác định và theo dõi vị trí của các bộ phận cơ thể chính, chẳng hạn như đầu, vai, khuỷu tay, cổ tay, hông, đầu gối và mắt cá chân, lập bản đồ hiệu quả bộ xương của một người.



## 1.2 Ứng dụng

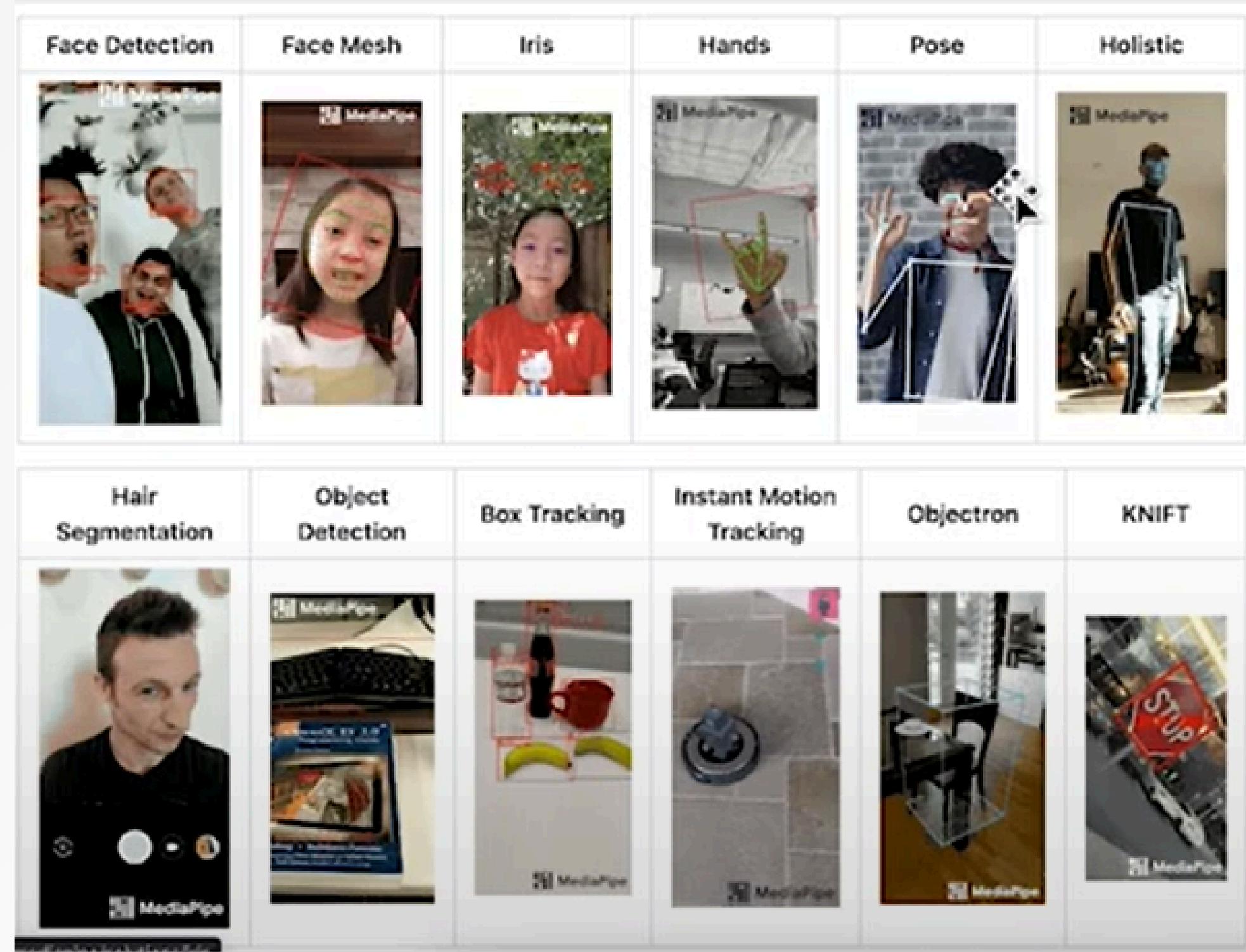
- Pose Estimation có nhiều ứng dụng trong các lĩnh vực khác nhau:
  - Thể thao: Giúp phân tích kỹ thuật và chuyển động của vận động viên. Dựa trên phân tích tư thế và cử động, các huấn luyện viên có thể phát hiện và giảm thiểu nguy cơ chấn thương bằng cách điều chỉnh kỹ thuật.
  - An ninh: Dùng trong giám sát an ninh để phát hiện các hành động đáng ngờ hoặc nguy hiểm.
  - Y tế: Pose Estimation có thể hỗ trợ trong việc chẩn đoán các vấn đề liên quan đến xương khớp hoặc cơ bắp bằng cách phân tích tư thế và chuyển động của bệnh nhân.

## 2.Tổng quan bài toán



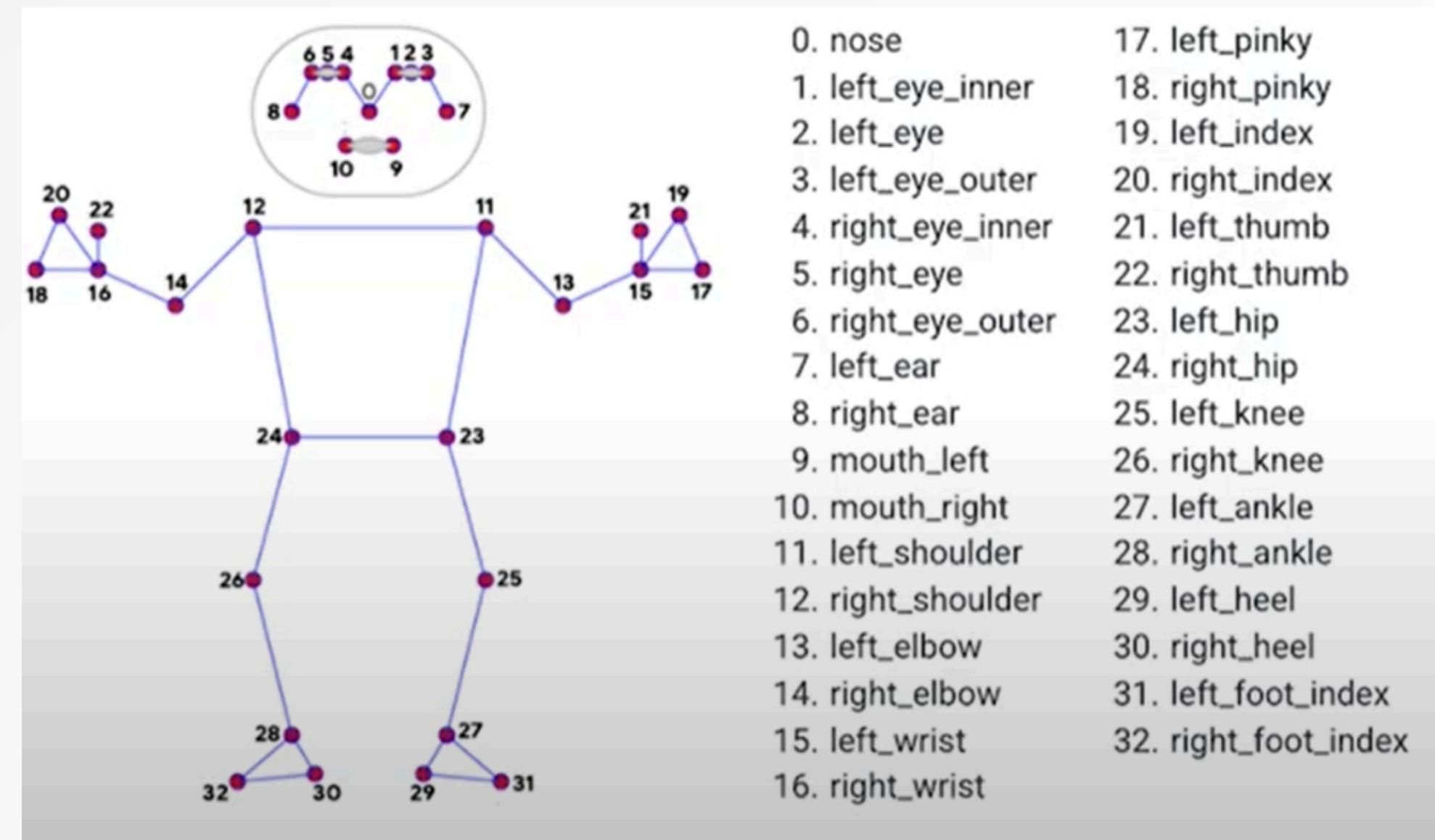
# 3. Media Pipe

MediaPipe là một khung phần mềm mã nguồn mở được phát triển bởi Google, cung cấp các giải pháp máy học sử dụng trong nhiều ứng dụng thời gian thực như nhận dạng khuôn mặt, phát hiện bàn tay, ước lượng tư thế, ... MediaPipe được thiết kế để hoạt động hiệu quả trên nhiều nền tảng như điện thoại di động, web, và máy tính để bàn, với khả năng chạy các mô hình ML trong thời gian thực.



# Media Pipe Pose

Pose Estimation là một trong những ứng dụng nổi bật mà MediaPipe hỗ trợ rất mạnh mẽ. MediaPipe Pose là một giải pháp cho việc ước lượng tư thế của con người trong thời gian thực, xác định vị trí của các điểm cơ thể chính (keypoints) trong hình ảnh hoặc video. MediaPipe Pose hỗ trợ nhận diện 33 điểm chính trên cơ thể, bao gồm các điểm trên mặt, tay, chân và thân mình.



# Media Pipe Pose

- Mỗi động tác sẽ được chia vào 1 file txt
- Số dòng sẽ là số frame
- Số cột là x,y,z,visibility của mỗi điểm trên khung xương (4 x 33)

```
≡ HANDSWING.txt > □ data
1 ,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,
2 0,0.6255581974983215,0.5930126905441284,-1.1593189239501953,0.9998264908790588,0.6479546427726746,0.5151227712631226,
3 1,0.6050856113433838,0.5913976430892944,-0.7987548112869263,0.9998430013656616,0.6368270516395569,0.5132938027381897,
4 2,0.6052550077438354,0.5911230444908142,-0.9286314249038696,0.9998499155044556,0.635761559009552,0.5133263468742371,-
5 3,0.6081809997558594,0.591285765171051,-0.8707637786865234,0.999864399433136,0.6366738080978394,0.5153220891952515,-0
6 4,0.6093840003013611,0.5929651856422424,-1.1152338981628418,0.9998753666877747,0.6369771361351013,0.5187601447105408,
7 5,0.6033887267112732,0.596572995185852,-1.1415860652923584,0.9998765587806702,0.6333593130111694,0.5224107503890991,-
8 6,0.6039913296699524,0.5963147282600403,-1.1498805284500122,0.999883234500885,0.6330262422561646,0.5221273303031921,-
9 7,0.6045209169387817,0.5942361950874329,-1.1159162521362305,0.9998876452445984,0.6332262754440308,0.5201348066329956,
10 8,0.6022700071334839,0.5948331356048584,-1.1138670444488525,0.9998906850814819,0.6317805051803589,0.5213520526885986,
11 9,0.5995835065841675,0.5949016809463501,-1.0915701389312744 Col 5: 3 0.34404754639,0.629934549331665,0.5269628763198853,
12 10,0.5976360440254211,0.5983773469924927,-0.983748435974121 0.9999064207077026,0.631373941898346,0.5281165242195129,-1
13 11,0.5994481444358826,0.599523663520813,-1.329690933227539,0.9999064207077026,0.631373941898346,0.5281165242195129,-1
14 12,0.6044996976852417,0.5996893048286438,-1.4545620679855347,0.9999117255210876,0.6326156854629517,0.5280492305755615
15 13,0.6070230007171631,0.5995514392852783,-1.417487382888794,0.9999154806137085,0.6328039169311523,0.527353823184967,-
16 14,0.607530951499939,0.5995975136756897,-1.595976710319519,0.9999169111251831,0.6328517198562622,0.5271537899971008,-
17 15,0.6066021919250488,0.5972887277603149,-0.9678305387496948,0.9999148845672607,0.6321051120758057,0.5243600606918335
18 16,0.6017225980758667,0.5931147933006287,-0.8960994482040405,0.9998654723167419,0.6307135820388794,0.5207358598709106
19 17,0.5995914340019226,0.5852864384651184,-1.0195825099945068,0.9998424649238586,0.6322355270385742,0.5158379077911377
20 18,0.5989285111427307,0.5875611901283264,-1.5019232034683228,0.9997870326042175,0.6325413584709167,0.5174306631088257
21 19,0.6019254922866821,0.6026638150215149,-1.1112207174301147,0.9997575283050537,0.6340181231498718,0.5292479395866394
22 20,0.6204736828804016,0.6106798648834229,-1.073803186416626,0.9996412992477417,0.6459107995033264,0.5381953120231628,
23 21,0.6265693306922913,0.6104109287261963,-0.8906135559082031,0.9994191527366638,0.6477329134941101,0.537962019443512,
24 22,0.6376644372940063,0.6150228381156921,-0.8943576812744141,0.9990820288658142,0.6550390720367432,0.542863667011261,
25 23,0.6457872986793518,0.6135985851287842,-0.9493484497070312,0.9991618394851685,0.6584648489952087,0.5389649271965027
```

# 4. Vấn đề bài toán

Pose Estimation có data là 1 chuỗi các hình ảnh bao gồm nhiều tư thế liên tục -> Mạng neural thông thường sẽ không có đủ khả năng để xử lý bài toán này -> LSTM

1. Xử lý chuỗi thời gian hiệu quả: LSTM có khả năng giữ lại thông tin quan trọng từ các khung hình trước đó và sử dụng chúng để cải thiện độ chính xác của các dự đoán trong khung hình hiện tại
2. Khả năng ghi nhớ dài hạn: LSTM có khả năng ghi nhớ dài hạn tốt hơn so với các mô hình RNN cơ bản
3. Giảm thiểu vấn đề vanishing gradient: So với RNN truyền thống, LSTM giải quyết tốt hơn vấn đề vanishing gradient. Cho phép thông tin quan trọng được lưu giữ và lan truyền qua nhiều lớp mạng mà không bị suy giảm.

# 5. Model

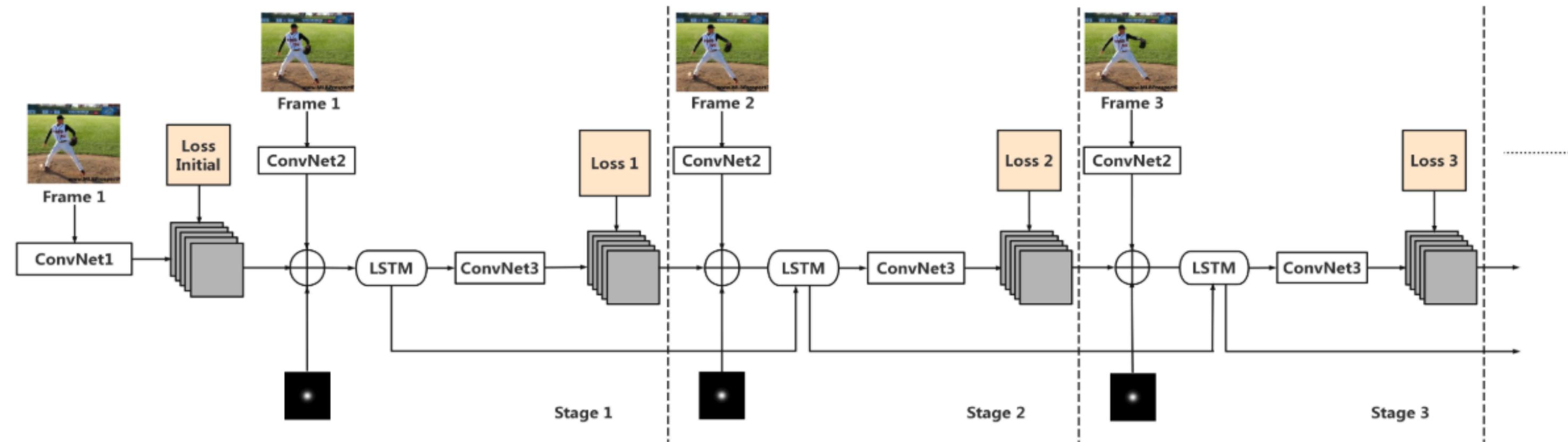


Figure 2. Network architecture for LSTM Pose Machines. This network consists of  $T$  stages, where  $T$  is the number of frames. In each stage, one frame from a sequence will be sent into the network as input. *ConvNet2* is a multi-layer CNN network for extracting features while an additional *ConvNet1* will be used in the first stage for initialization. Results from the last stage will be concatenated with newly processed inputs plus a central Gaussian map, and they will be sent into the *LSTM* module. Outputs from *LSTM* will pass *ConvNet3* and produce predictions for each frame. The architectures of those *ConvNets* are the same as the counterparts used in the CPM model [36] but their weights are shared across stages. *LSTM* also enables weight sharing, which reduces the number of parameters in our network.

# 5. Model

## 5.1 Formular

$$\mathbf{b}_t = g(\tilde{\mathcal{L}}(\mathcal{F}'(X_t))), \quad t = 1,$$

$$\mathbf{b}_t = g(\tilde{\mathcal{L}}(\mathcal{F}(X_t) \oplus \mathbf{b}_{t-1}))), \quad t = 2, 3, \dots, T.$$

$\mathcal{F}'(X_t)$  can be decomposed as  $\mathcal{F}_0(X_t) \oplus \mathcal{F}(X_t)$

# 5. Model

## 5.2 Operation

$X(t)$ : input at t

$h(t-1)$ : output at t-1

$W$ : corresponding weight matrix

$\epsilon$ : bias

$\sigma$ : activation function

$C(t)$ : memory cell

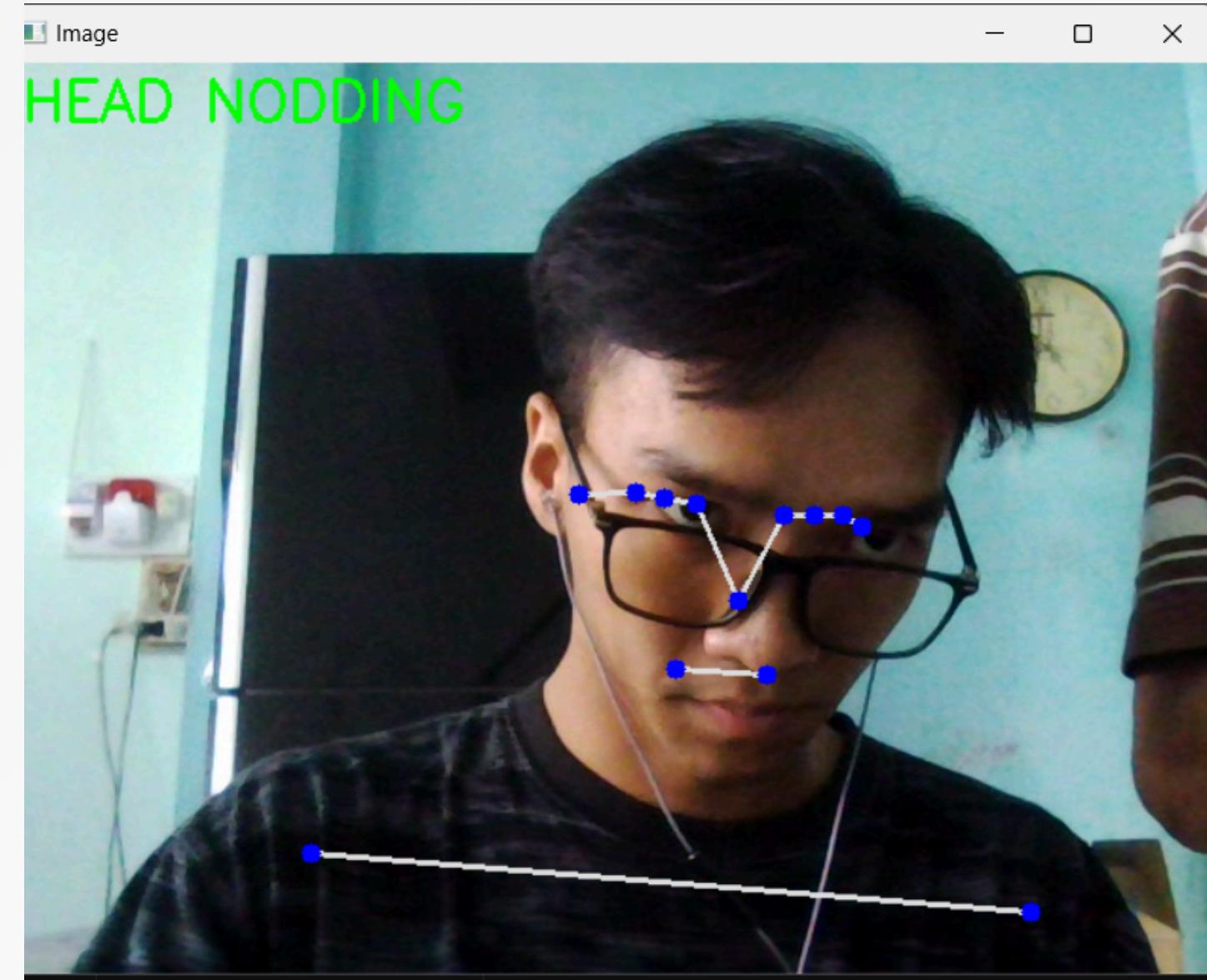
$h(t)$ : hidden state

$$\begin{aligned}g_t &= \varphi(W_{xg} * X_t + W_{hg} * h_{t-1} + \epsilon_g), \\i_t &= \sigma(W_{xi} * X_t + W_{hi} * h_{t-1} + \epsilon_i), \\f_t &= \sigma(W_{xf} * X_t + W_{hf} * h_{t-1} + \epsilon_f), \\o_t &= \sigma(W_{xo} * X_t + W_{ho} * h_{t-1} + \epsilon_o), \\C_t &= f_t \odot C_{t-1} + i_t \odot g_t, \\h_t &= o_t \odot \varphi(C_t)\end{aligned}$$

# 5. Model

- Loss: categorical crossentropy
- Lớp Dense sau cùng đưa ra 4 classes, và hàm kích hoạt là soft-max
- Số epoch: 20
- Hàm tối ưu: Adam

# 6. Result



Kết quả nhận diện được khá chính xác với 4 bộ data tương ứng với 4 động tác

# 6. Result

```
Epoch 15/20
60/60 —————— 1s 10ms/step - accuracy: 0.9861 - loss: 0.0407 - val_accuracy: 1.0000 - val_loss: 0.0019
Epoch 16/20
60/60 —————— 1s 10ms/step - accuracy: 0.9973 - loss: 0.0084 - val_accuracy: 0.9810 - val_loss: 0.0620
Epoch 17/20
60/60 —————— 1s 9ms/step - accuracy: 0.9851 - loss: 0.0466 - val_accuracy: 0.9979 - val_loss: 0.0125
Epoch 18/20
60/60 —————— 1s 9ms/step - accuracy: 0.9943 - loss: 0.0201 - val_accuracy: 0.9937 - val_loss: 0.0098
Epoch 19/20
60/60 —————— 1s 14ms/step - accuracy: 0.9906 - loss: 0.0339 - val_accuracy: 0.9979 - val_loss: 0.0141
Epoch 20/20
60/60 —————— 1s 16ms/step - accuracy: 0.9982 - loss: 0.0136 - val_accuracy: 0.9979 - val_loss: 0.0057
```

# 7. Limitation

- Vẫn bị trùng lặp giữa những động tác gần giống nhau
- Xử lý với 2 người trở lên trong cùng 1 khung ảnh



THANK YOU