ỨNG DỤNG PHÁT HIỆN ĐIỂM NEO TRÊN CƠ THỂ

BÁO CÁO ĐỒ ÁN MÔN XỬ LÝ ẢNH VIDEO SỐ

19127120 – Ngô Nhật Du

19127395 – Phan Đức Hiển

19127396 – Phan Thiên Vinh Hiển

HCMUS - Nhóm 3

1. HUMAN POSE ESTIMATION (HPE): ỨNG DỤNG PHÁT HIỆN ĐIỂM NEO TRÊN CƠ THỂ

1.1. Khoa Học

HPE là một công nghệ dựa trên thị giác máy tính để xác định và phân tích để hiểu thông tin hình học và chuyển động trên các khớp nối của cơ thể con người.

HPE về cơ bản là một cách để nắm bắt một tập hợp các tọa độ cho mỗi khớp, được gọi là điểm chính có thể mô tả tư thế của một người. Các mối kết nối giữa các điểm này là một cặp kết nối, không phải tất cả các điểm đều có thể tạo thành một cặp.

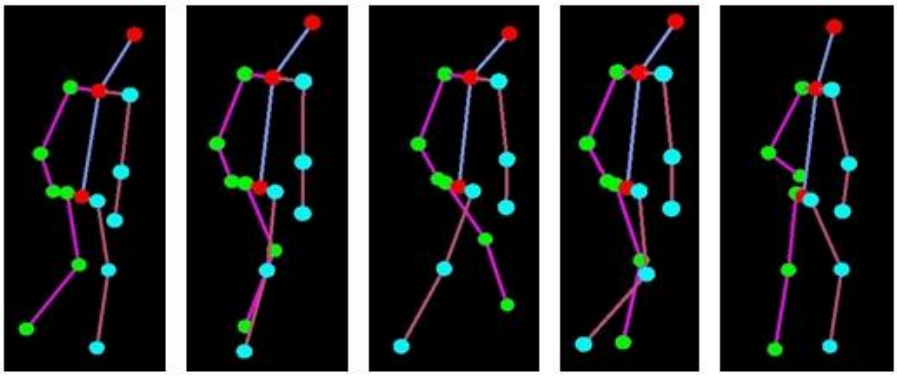
Mục đính của HPE là tạo dựng lại bộ xương người và được xử lý cho các ứng dụng cụ thể.

1.2. Ứng Dụng

1.2.1. **Nhận biết hoạt động** (Activity Recognition)

Theo dõi sự thay đổi trong tư thế của một người trong một khoảng thời gian cũng có thể được sử dụng để nhận dạng hoạt động, cử chỉ và dáng đi. Có một số trường hợp sử dụng cho cùng một trường hợp, bao gồm:

- Ứng dụng để phát hiện xem một người đã ngã xuống hoặc bị bệnh.
- Ứng dụng có thể tự động dạy các chế độ tập luyện phù hợp, các kỹ thuật thể thao và các hoạt động khiêu vũ.
- Ứng dụng có thể hiểu ngôn ngữ ký hiệu toàn thân. (Ví dụ: Tín hiệu đường băng sân bay, tín hiệu cảnh sát giao thông, v.v.).
- Ứng dụng có thể tăng cường bảo mật và giám sát.



Hình 1 Theo dõi dáng đi của người cho mục đích an ninh.

1.2.2. **Chụp chuyển động** (Motion Capture) **và Thực tế tăng cường** (Augmented Reality)

Ứng dụng công nghệ mô phỏng hình ảnh bằng máy tính (Computer-generated imagery): Đồ họa, phong cách, cải tiến ưa thích, thiết bị và tác phẩm nghệ thuật có thể tái tạo lại hình ảnh con người người nếu có thể ước lượng được các điểm neo trên cơ thể. Bằng cách theo dõi các thay đổi trong chuyển động của con người, đồ họa được kết xuất có thể kết hợp "vừa vặn" một cách tự nhiên với người khi họ di chuyển.

Hình 2 Ví dụ về kết xuất CGI.

Có thể thấy một ví dụ trực quan tốt về những gì có thể thực hiện được qua **Animoji**. Mặc dù phần trên chỉ theo dõi cấu trúc của một khuôn mặt, nhưng ý tưởng này có thể được ngoại suy cho các điểm chính của một người. Các khái niệm tương tự có thể được tận dụng để hiển thị các yếu tố Thực tế tăng cường (AR) có thể bắt chước các chuyển động của một người.

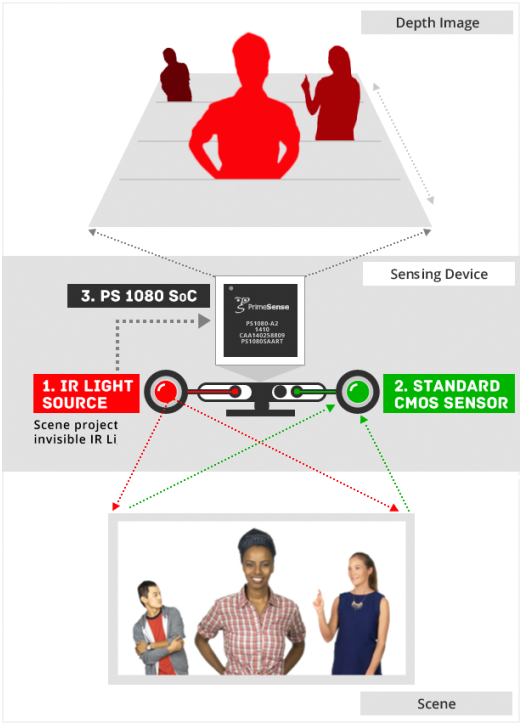
Hình 3 Ví dụ về Animoji

1.2.3. Huấn luyện Robot

Thay vì lập trình thủ công cho rô-bốt đi theo quỹ đạo (follow trajectories), rô-bốt có thể được chế tạo để tự đi theo quỹ đạo của điểm neo của bộ xương người đang thực hiện một hành động. Một người hướng dẫn con người có thể dạy một cách hiệu quả các hành động nhất định của robot bằng cách chỉ cần trình diễn những thao tác tương tự. Sau đó, robot có thể tính toán cách di chuyển các khớp nối của nó để thực hiện cùng một hành động.

1.2.4. Theo dõi chuyển động (Motion Tracking) cho bảng điều khiển (consoles)

Ứng dụng để theo dõi chuyển động của đối tượng con người để chơi trò chơi tương tác (interactive gaming). Một trong những game phổ biến là Kinect đã sử dụng ước tính tư thế 3D (sử dụng dữ liệu cảm biến hồng ngoại) để theo dõi chuyển động của người chơi và sử dụng nó để hiển thị hành động của các nhân vật ảo.



Hình 4 Ví dụ về cảm biến Kinect

2. PHÁT BIỂU BÀI TOÁN

2.1. Đầu Vào

Hình ảnh (ở định dạng opencv) của tư thế con người sẽ được ước tính.
Hình ảnh có thể là:

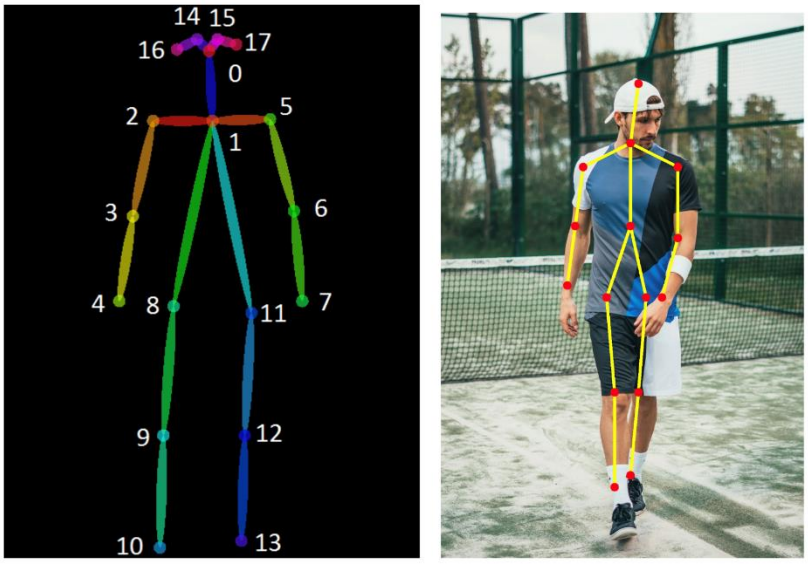
- một hình ảnh duy nhất có chiều cao, chiều rộng, kênh màu BGR.
- một chồng n hình ảnh có n, chiều cao, chiều rộng, kênh màu BGR.

2.2. Đầu Ra

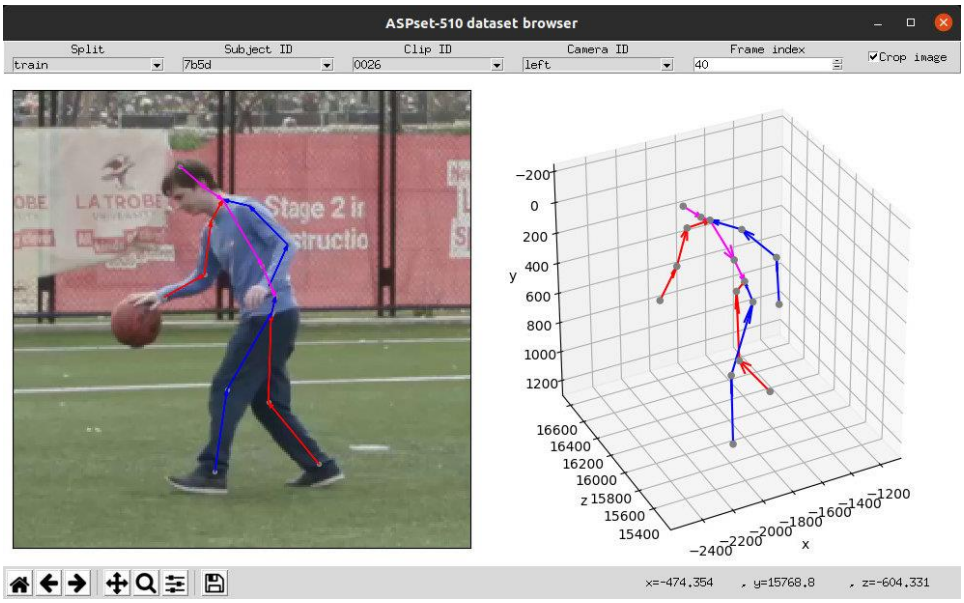
- Hình ảnh/ video kèm với bộ xương người.
- một mảng phức tạp chứa khớp người cho mỗi người (được phát hiện)
    - nếu là một hình ảnh duy nhất trả về dạng có có số người, số khớp và 3 kênh màu

- nếu là một chồng n hình ảnh thì trả về danh sách mảng có số người, số khớp và 3 kênh màu

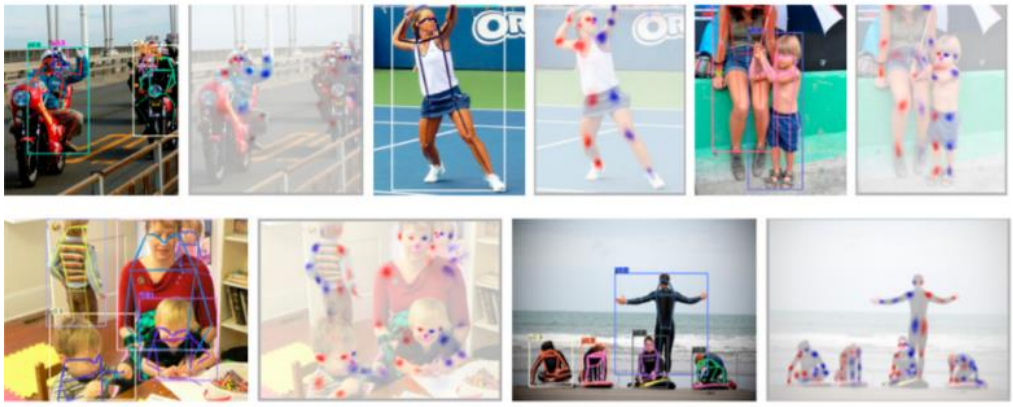
```
return_heatmaps, the class returns a list with (heatmaps, human joints)
return_bounding_boxes, the class returns a list with (bounding boxes, human joints)
return_heatmaps and return_bounding_boxes, the class returns a list with (heatmaps,
bounding boxes, human joints)
```



Hình 5 Thứ tự các điểm chính (keypoints) được đánh dấu



Hình 6 Hướng vector của các diểm chính (keypoints)



Hình 7 Đầu ra của heatmaps, bounding boxes và tags.

2.3. Cài Đặt

Library	Module	Framework
numpy>=1.16 ffmpeg-python>=0.2.0 matplotlib>=3.0.2 opencv-python>=3.4 Pillow>=5.4 torch>=1.4.0 torchvision>=0.5.0 tqdm>=4.26	munkres>=1.1.2	vidgear>=0.1.4

2.4. Thách Thức

- Thách thức trong bài toán được đặt ra cho việc dự đoán tư thế chính xác cho người nhỏ con do những thách thức về sự thay đổi quy mô, tức là nâng cao hiệu suất của những người nhỏ mà không phải hy sinh hiệu suất của những người lớn. Cần tạo ra các bản đồ nhiệt nhận biết về quy mô và chính xác hơn về mặt không gian để dự đoán điểm mấu chốt từ dưới lên một cách tự nhiên và đơn giản mà không hy sinh chi phí tính toán.
- Thứ hai, thách thức bài toán ở mức độ đông người

3. NGHIÊN CỨU LIÊN QUAN

	Phương Pháp	Độ hiệu quả	Ưu điểm	Khuyết điểm
1	DeepPose	60%	Khi một số khớp nhất định bị ẩn, chúng có thể được ước tính nếu tư thế được lý luận một cách tổng thể.	Việc rút lui đến các vị trí XY rất khó và làm tăng thêm độ phức tạp trong việc học, điều này làm suy yếu khả năng tổng quát và do đó hoạt động kém hơn ở một số vùng nhất định.
2	Efficient Object Localization using Convolutional Networks	67%	Khôi phục độ chính xác không gian bị mất do gộp trong mô hình ban đầu.	Thiếu mô hình cấu trúc. Việc lập mô hình cấu trúc này sẽ giúp bạn dễ dàng xác định các điểm quan trọng có thể nhìn thấy và giúp bạn có thể ước tính các điểm chính bị tắc.
3	Convolutional Pose Machines	78%	Cung cấp một khung dự đoán tuần tự để học các mô hình không gian tiềm ẩn phong phú và hoạt động rất tốt cho tư thế con người.	
4	Human Pose Estimation with Iterative Error Feedback	81.3%	Dự đoán sai sót của các ước tính hiện tại và sửa chúng lặp đi lặp lại.	
5	Stacked Hourglass Networks for Human Pose Estimation	90.9%	- Các bước gộp và lấy mẫu các lớp trông giống như một chiếc đồng hồ cát và những lớp này được xếp chồng lên nhau. Thiết kế của mạng đồng hồ cát được thúc đẩy bởi nhu cầu nắm bắt thông tin ở mọi quy mô. - Mạng đồng hồ cát nắm bắt thông tin ở mọi tỷ lệ. Bằng cách này, thông tin toàn cầu và địa phương được thu thập hoàn chỉnh và được mạng lưới sử dụng để tìm hiểu các dự đoán.	Rất phức tạp.
6	Simple Baselines for Human Pose Estimation and Tracking	79%	- Cấu trúc mạng khá đơn giản và bao gồm ResNet + một số lớp deconvolutions ở cuối.	Không có hiệu quả cao.
7	Higher High-Resolution Net	70.5% 67.6% (CrowdPose)	Vượt trội hơn tất cả các phương pháp từ dưới lên hiện có về ước tính tư thế con người.	

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Bottom-up: keypoint detection and grouping										
OpenPose [6]	–	–	–	–	61.8	84.9	67.5	57.1	68.2	66.5
Associative Embedding [39]	–	–	–	–	65.5	86.8	72.3	60.6	72.6	70.2
PersonLab [46]	–	–	–	–	68.7	89.0	75.4	64.1	75.5	75.4
MultiPoseNet [33]	–	–	–	–	69.6	86.3	76.6	65.0	76.3	73.5
Top-down: human detection and single-person keypoint detection										
Mask-RCNN [21]	ResNet-50-FPN	–	–	–	63.1	87.3	68.7	57.8	71.4	–
G-RMI [47]	ResNet-101	353 × 257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression [60]	ResNet-101	256 × 256	45.0M	11.0	67.8	88.2	74.8	63.9	74.0	–
G-RMI + extra data [47]	ResNet-101	353 × 257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3
CPN [11]	ResNet-Inception	384 × 288	–	–	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [17]	PyraNet [77]	320 × 256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	–
CFN [25]	–	–	–	–	72.6	86.1	69.7	78.3	64.1	–
CPN (ensemble) [11]	ResNet-Inception	384 × 288	–	–	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [72]	ResNet-152	384 × 288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W32	HRNet-W32	384 × 288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet-W48	HRNet-W48	384 × 288	63.6M	32.9	**75.5**	**92.5**	**83.3**	**71.9**	**81.5**	**80.5**
HRNet-W48 + extra data	HRNet-W48	384 × 288	63.6M	32.9	**77.0**	**92.7**	**84.5**	**73.4**	**83.1**	**82.0**

Hình 8 Kết quả so sánh giữa độ chính xác trên các lớp dữ liệu khác nhau

4. PHƯƠNG PHÁP NGHIÊN CỨU VÀ TIẾN HÀNH

4.1. Nguyên Lý

3 cách mô hình hóa cơ thể người:

4.1.1. Skeleton-based model (mô hình dựa trên bộ xương)

Tập hợp các khớp như cổ, vai, khuỷu tay, cổ tay, hông, khuỷu chân, mắt cá chân và các hướng chi bao gồm cấu trúc xương của cơ thể người.

Tính linh hoạt cao, được sử dụng trong cả 2D và 3D HPE.

4.1.2. Contour-based model (mô hình dựa trên đường viền)

Gồm đường viền, chiều rộng của cơ thể người và các chi.

4.1.3. Volume-based model (mô hình dựa trên khối lượng)

Gồm hình dạng và tư thế cơ thể người, thể hiện bằng các mô hình dựa trên khối lượng có lưới và hình dạng hình học.

4.2. Phương Pháp

4.2.1. Phương pháp Tiếp cận Cổ Điển:

- Ý tưởng cơ bản ở đây là đại diện cho một đối tượng bởi một tập “các bộ phận” được sắp xếp theo cấu hình biến dạng. Một "bộ phận" là một mẫu xuất hiện được khớp trong một hình ảnh.
- Khi các bộ phận được tham số hóa theo vị trí và hướng pixel, kết quả của cấu trúc có thể mô hình hóa khớp nối phù hợp trong HPE.

Các phương pháp tiếp cận cổ điển thường đề cập đến các kỹ thuật và phương pháp liên quan đến các thuật toán học máy nuốt.

Ví dụ, những nghiên cứu trước đó về HPE bao gồm việc thực hiện Random Forest trong “khung cấu trúc hình ảnh”, được sử dụng để dự đoán các khớp trong cơ thể con người.

Khung cấu trúc hình ảnh (PSF) thường được coi là một trong những phương pháp truyền thống của HPE. PSF chứa hai thành phần:

- **Discriminator** (bộ phân biệt): Mô hình hóa khả năng một bộ phận nhất định tại một vị trí cụ thể, xác định các bộ phận cơ thể.
- **Prior** (ưu tiên): Mô hình hóa phân phối xác suất so với tư thế bằng cách sử dụng đầu ra của bộ phân biệt; tư thế mô hình phải thực tế.

Mục đích của PSF là đại diện cho cơ thể con người dưới dạng tập hợp các tọa độ cho từng bộ phận cơ thể trong một hình ảnh đầu vào nhất định. PSF sử dụng các bộ hồi quy chung phi tuyến, lý tưởng là một bộ hồi quy Random Forest hai lớp.

Các mô hình này hoạt động tốt khi hình ảnh đầu vào có các chi rõ ràng và có thể nhìn thấy được. Tuy nhiên, không thể chụp và lập mô hình các chi bị ẩn hoặc không nhìn thấy từ một góc nhất định.

Để khắc phục những vấn đề này, các phương pháp mô tả đặc trưng Histogram oriented Gaussian (HOG), đường viền, biểu đồ, v.v. đã được sử dụng. Tuy sử dụng các phương pháp hỗ trợ này, mô hình cổ điển thiếu độ chính xác, tính tương quan và khả năng tổng quát hóa.

4.2.2. Phương pháp Tiếp cận Deep-Learning:

Các phương pháp tiếp cận dựa trên học sâu được xác định rõ ràng nhờ khả năng tổng quát hóa bất kỳ chức năng nào (nếu có đủ dữ liệu).

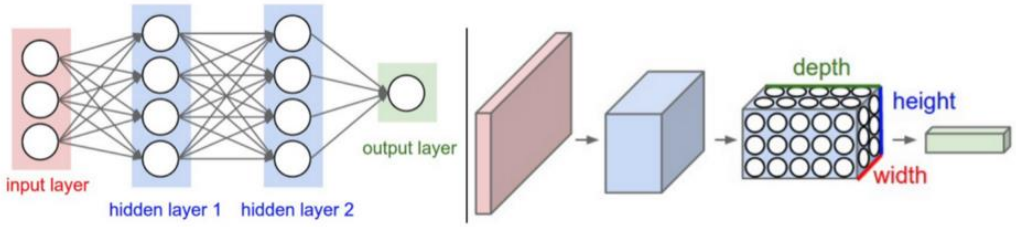
Khi nói đến các tác vụ của thị giác máy tính, mạng nơ-ron phức hợp sâu (deep convolutional neural networks) vượt qua tất cả các thuật toán khác, và điều này cũng đúng đối với HPE.

CNN có khả năng trích xuất các mẫu và biểu diễn từ hình ảnh đầu vào đã cho với độ chính xác cao hơn bất kỳ thuật toán nào khác; do đó CNN rất hữu ích cho các tác vụ như phân loại, phát hiện và phân đoạn.

Không giống như cách tiếp cận cổ điển, CNN có thể học các tính năng phức tạp khi được cung cấp đủ dữ liệu đào tạo.

Một mạng nơ-ron thông thường chứa ba lớp ẩn, trong khi mạng sâu có thể có tới 120-150. Cung cấp cho hệ thống máy tính rất nhiều dữ liệu, dữ liệu này có thể sử dụng để đưa ra quyết định về các dữ liệu khác, được cung cấp thông qua mạng nơ-ron.

Mạng học sâu có thể học các tính năng trực tiếp từ dữ liệu mà không cần trích xuất tính năng thủ công.



Convolution Layer – tính toán đầu ra của các vùng cục bộ của hình ảnh bằng cách sử dụng bộ lọc

Pooling Layer – thực hiện lấy mẫu xuống dọc theo các kích thước không gian

Fully-Connected-Layer – tính toán kết quả cuối cùng

DNN có lợi thế hơn trong ước tính tư thế một cá thể, bất lợi hơn đối với nhiều cá thể.

Các khó khăn thường gặp:

- Một hình ảnh có thể chứa nhiều cá thể ở các vị trí khác nhau.
- Khi số lượng cá thể tăng lên, sự tương tác giữa các bên tăng lên dẫn đến sự phức tạp trong tính toán.
- Sự gia tăng độ phức tạp trong tính toán thường dẫn đến tăng thời gian suy luận trong thời gian thực.

Các hướng giải quyết:

4.2.2.1. Top-down (từ trên xuống):

Mạng độ phân giải cao (HRNet) là một mạng thần kinh để ước tính tư thế con người. Nó là một kiến trúc được sử dụng trong các bài toán xử lý ảnh để tìm những gì chúng ta biết là các điểm chính (khớp) liên quan đến đối tượng hoặc người cụ thể trong ảnh. Một lợi thế của kiến trúc này so với các kiến trúc khác là hầu hết các phương pháp hiện có phù hợp với các biểu diễn tư thế có độ phân giải cao từ các biểu diễn có độ phân giải thấp đối với việc sử dụng các mạng có độ phân giải cao-thấp. Mạng nơ-ron duy trì các biểu diễn có độ phân giải cao khi ước tính các tư thế. Ví dụ: kiến trúc HRNet này rất hữu ích cho việc phát hiện tư thế của con người trong các môn thể thao được truyền hình.

High-Resolution Network:

HRNet là một thuật toán hiện đại trong lĩnh vực phân đoạn ngữ nghĩa, phát hiện điểm neo khuôn mặt và tư thế người. Nó đã cho thấy kết quả vượt trội trong việc phân đoạn ngữ nghĩa trên các tập dữ liệu như PASCAL Context, LIP, Cityscapes, AFLW, COFW và 300W.

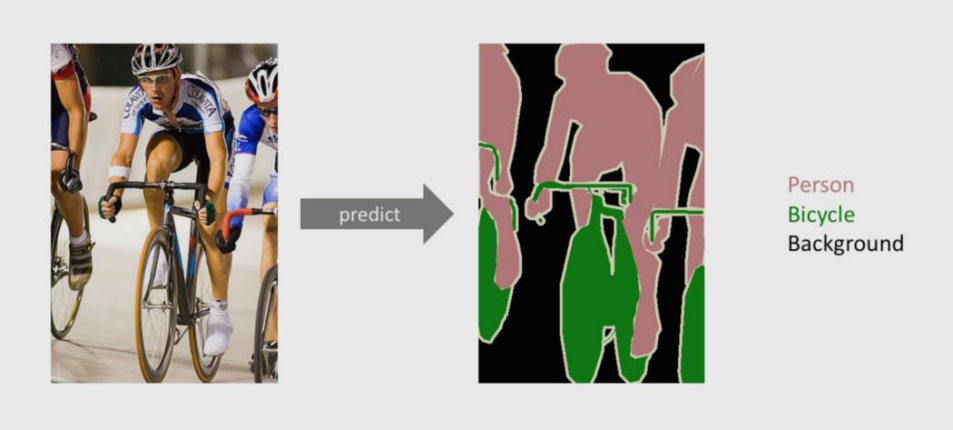
Ứng dụng:

Phân đoạn ngữ nghĩa:

Được sử dụng để phân loại cấu trúc của một hình ảnh thành các lớp nhất định. Điều này được thực hiện bằng cách gắn nhãn mỗi pixel với một lớp nhất định. Mục đích của phân đoạn hình ảnh là để thuật toán phân đoạn hình ảnh thành các lớp và do đó có cấu trúc nhất định.

Được ứng dụng trong lái xe tự động, chẩn đoán hình ảnh y tế, nhận dạng chữ viết tay.

Trong ví dụ dưới đây, tất cả các pixel đại diện cho người đi xe đạp là một người trong lớp và tất cả các pixel đại diện cho xe đạp là xe đạp của lớp.



Phát hiện điểm neo trên khuôn mặt và cơ thể

Được sử dụng để nhận dạng và xác định vị trí các vùng nhất định trên khuôn mặt như mũi, miệng, mắt hoặc lông mày. Trong hình ảnh sau, bạn có thể thấy rằng bằng cách sử dụng OpenCV, có thể phát hiện lông mày, mũi và miệng (có thể nhìn thấy qua các chấm đỏ trên hình ảnh bên trái).

Các bộ lọc hoán đổi khuôn mặt trong Snapchat hoặc Instagram hoạt động với tính năng phát hiện điểm neo khuôn mặt để phát hiện vị trí của một phần nhất định trên khuôn mặt của bạn. Hơn nữa, nó được sử dụng để ước lượng biến hình khuôn mặt và tư thế đầu.

Phát hiện điểm neo trên cơ thể người tương tự như phát hiện điểm neo khuôn mặt ngoại trừ nó được áp dụng cho toàn bộ cơ thể và nó liên quan nhiều hơn đến chuyển động. Thay vì các vùng trên khuôn mặt, nó phát hiện các điểm chính về ngữ nghĩa như trái phải, đầu gối phải, v.v.

4.2.2.2. Bottom-up (từ dưới lên):

Xác định vị trí cá thể người trong hình ảnh hoặc video, sau đó ước tính các bộ phận, sau đó tính toán tư thế.

OpenPose:

OpenPose là thư viện phát hiện tư thế người nhiều người trong thời gian thực, lần đầu tiên cho thấy khả năng cùng phát hiện các điểm chính trên cơ thể người, bàn chân, bàn tay và khuôn mặt trên các hình ảnh đơn lẻ. OpenPose có khả năng phát hiện tổng cộng 135 điểm chính.

OpenPose là tính năng phát hiện nhiều người theo thời gian thực mã nguồn mở, với độ chính xác cao trong việc phát hiện các điểm chính trên cơ thể, bàn chân, bàn tay và khuôn mặt. Một lợi thế của OpenPose là nó là một API cho phép người dùng linh hoạt trong việc lựa chọn hình ảnh nguồn từ các trường camera, webcam và những thứ khác, quan trọng hơn là đối với các ứng dụng hệ thống nhúng (ví dụ: tích hợp với hệ thống và camera CCTV). Nó hỗ trợ các kiến trúc phần cứng khác nhau, chẳng hạn như GPU CUDA, GPU OpenCL hoặc các thiết bị chỉ dành cho CPU. Phiên bản nhẹ đủ hiệu quả cho các ứng dụng suy luận Edge với xử lý trên thiết bị trong thời gian thực với các thiết bị cạnh.

Tính năng của thư viện phát hiện điểm neo trên bộ xương người OpenPose:

1. Phát hiện điểm chính (keypoint) 3D một người theo thời gian thực và 2D nhiều người trong thời gian thực.
2. Theo dõi một cá thể để tăng tốc độ phát hiện và làm mịn hình ảnh.
3. Công cụ hiệu chuẩn để ước tính các thông số camera bên ngoài, bên trong và độ biến thể.

4.2.2.3. So sánh

OpenPose bỏ qua các giá trị bị thiếu khi các khớp không nhìn thấy được, trong khi HRNet cố gắng ước tính vị trí của các khớp không nhìn thấy. Nó có thể không được nhìn thấy rõ ràng trong các video trên, nhưng có thể nhận thấy được điều này khi chúng ta nhìn từ từ từng khung hình trong video thứ 2 (khi người phụ nữ thực hiện các động tác xoay người).

Độ chính xác: Về độ chính xác, cả hai kiểu máy đều hoạt động tốt tương tự nhau ở 2 video đầu tiên, nhưng hiệu suất giảm ở video thứ 3. Điều này có thể do không có đủ hình ảnh về tư thế không thẳng đứng trong dữ liệu đào tạo COCO. Theo bài báo của HRNet, mô hình tốt nhất của họ đạt được mAP là 77,0 so với mAP là 61,8 mà OpenPose đạt được trên tập dữ liệu COCO. Vì vậy, HRNet là người chiến thắng về độ chính xác (mAP cao hơn 24,5%).

4.3. Giải Thuật

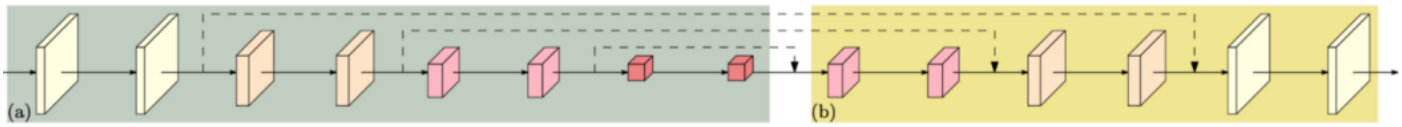
Kiến trúc HRNet:

Tất cả các ứng dụng và phương pháp này đều dựa trên Convolutional Neural Networks, là nền tảng cơ bản của HRNet.

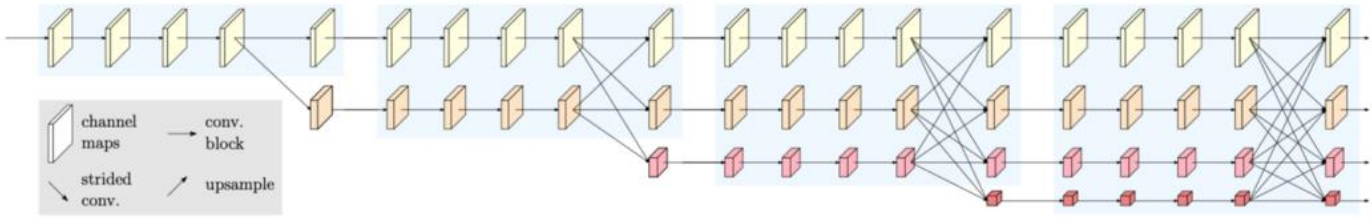
Các biểu diễn có độ phân giải cao mạnh mẽ đóng một vai trò thiết yếu trong các vấn đề gắn nhãn pixel và vùng, ví dụ: phân đoạn ngữ nghĩa, ước tính tư thế người, phát hiện mốc khuôn mặt và phát hiện đối tượng. Với số lượng pixel ngày càng tăng và nhiều vấn đề về video hơn, độ phân giải cao có thể đóng một vai trò ngày càng tăng trong tương lai.

Mô hình HRNet (Mạng độ phân giải cao) đã vượt trội hơn tất cả các phương pháp hiện có về các nhiệm vụ Phát hiện điểm chính, Ước tính tư thế nhiều người và Ước tính tư thế trong tập dữ liệu COCO và là phương pháp gần đây nhất.

HRNet thực hiện theo một ý tưởng rất đơn giản. Hầu hết các bài báo trước đây đều đi từ đại diện độ phân giải cao đến thấp đến cao. HRNet duy trì biểu diễn có độ phân giải cao trong toàn bộ quá trình và điều này hoạt động rất tốt.

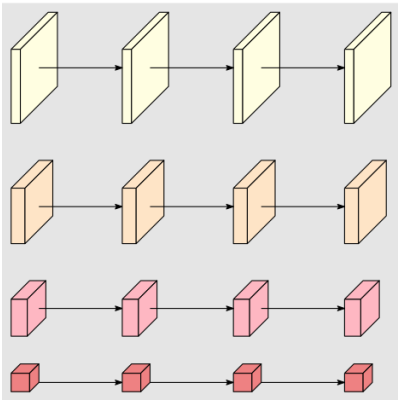


Hình 9 Cấu trúc khôi phục độ phân giải cao từ độ phân giải thấp.



Hình 10 Kiến trúc của giải thuật HRNet

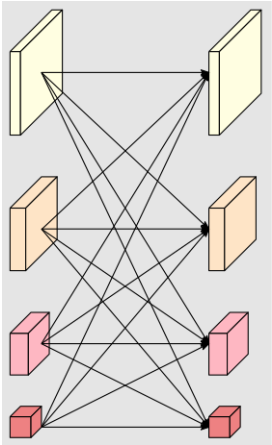
Cấu trúc của thuật toán HRNet được thể hiện trong Hình 6 có khả năng duy trì các biểu diễn có độ phân giải cao trong toàn mạng. Các giai đoạn sau được hình thành bằng cách thêm dần các mạng con có độ phân giải cao đến độ phân giải thấp từ các mạng con độ phân giải cao như giai đoạn đầu tiên và kết nối song song các mạng con đa độ phân giải. Trong suốt quá trình, phản ứng tổng hợp lặp đi lặp lại nhiều tỷ lệ được thực hiện bằng cách tương tác lặp lại thông tin trên các mạng con đa độ phân giải song song, để mỗi biểu diễn có độ phân giải cao đến độ phân giải thấp nhận được thông tin.



Mỗi khối đại diện cho một khối đa độ phân giải như đã nói ở trên, một khối kết nối song song “các khối có độ phân giải cao đến thấp”. Quá trình xử lý song song được thể hiện bằng nhiều đường bản đồ kênh (channel-map-lines) bên dưới nhau.

Bản đồ kênh màu vàng thể hiện độ phân giải cao nhất, bản đồ kênh nhỏ màu đỏ thể hiện độ phân giải thấp nhất.

Khối thứ tư xử lý song song 4 độ phân giải.

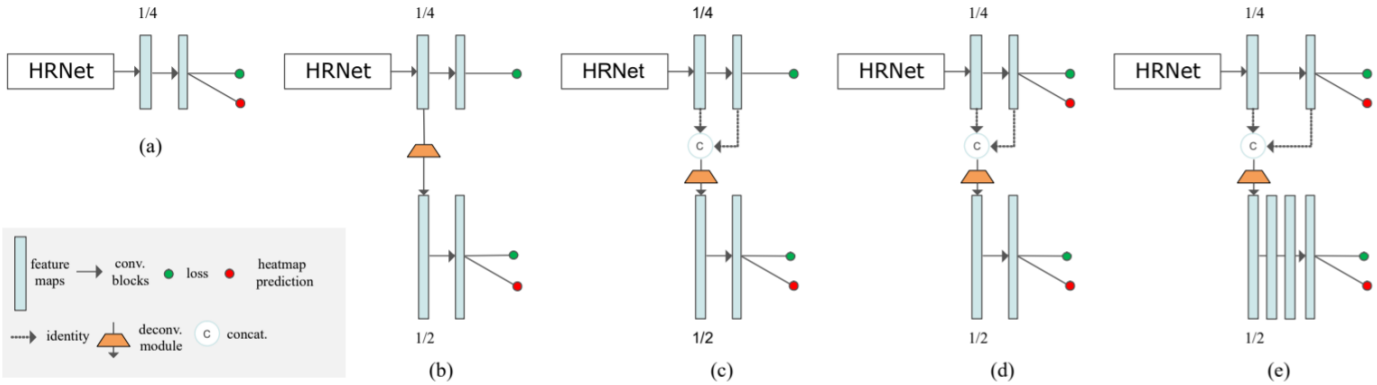


Vào cuối mỗi giai đoạn, có thể thấy kết nối đầy đủ đến nhóm đa độ phân giải của giai đoạn tiếp theo, trông giống như hình ảnh ở bên trên. Phần này của HRNet được gọi là tích chập đa độ phân giải (multi-resolution convolution).

Training details

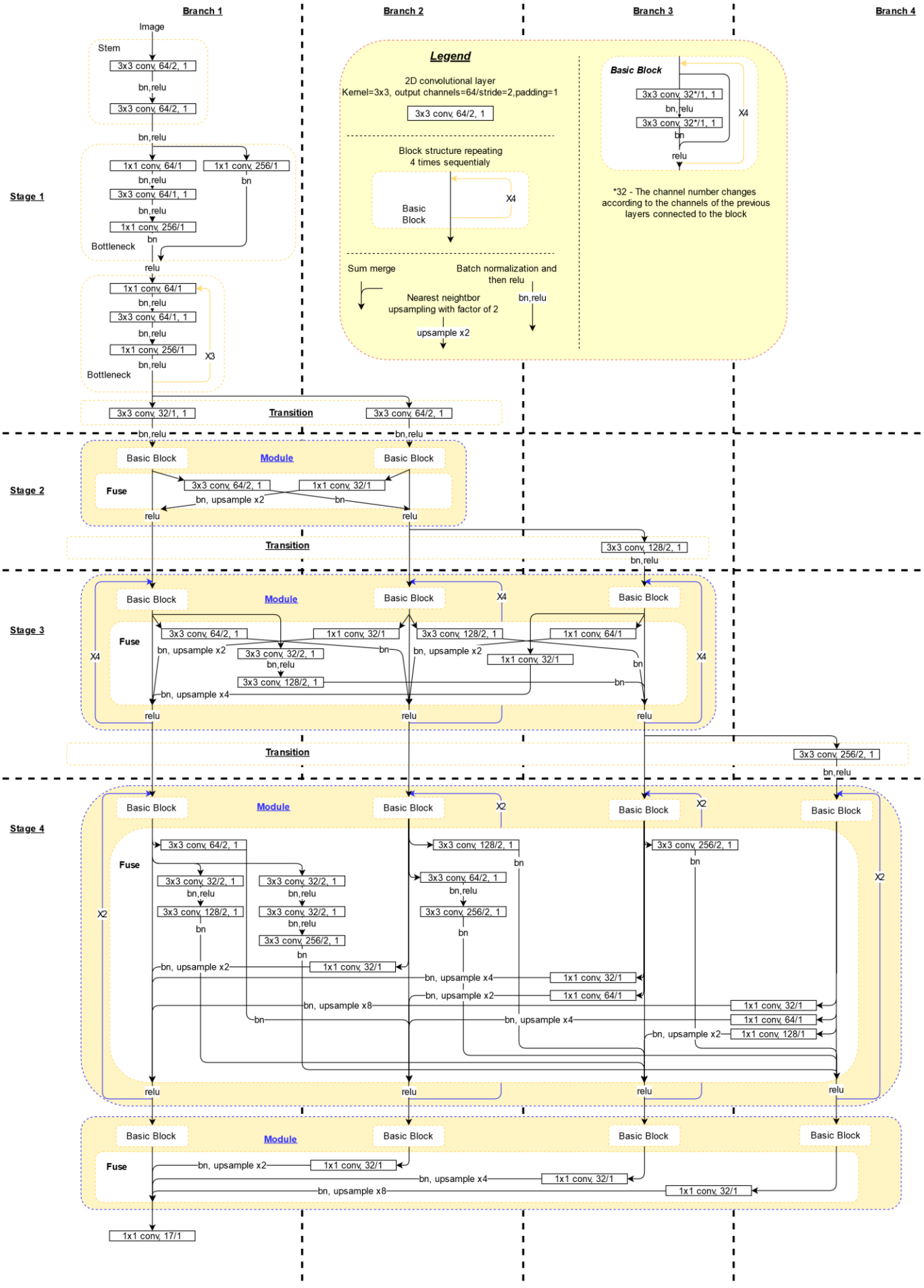
- COCO - hơn 200 nghìn hình ảnh và 250 nghìn cá thể người được gắn nhãn với 17 điểm chính. Đánh giá tập dữ liệu COCO cũng yêu cầu đánh giá các hộp giới hạn người, điều này được thực hiện bằng mạng FasterRCNN. Chỉ số đánh giá là độ tương đồng của điểm chính của đối tượng (OKS) - một chỉ số tiêu chuẩn về độ chính xác của việc phát hiện điểm chính.
- MPII - khoảng 25K hình ảnh với 40K đối tượng được gắn nhãn với 17 điểm chính. Đánh giá MPII được thực hiện với các hộp giới hạn có chú thích từ tập dữ liệu.

Overall Framework:



Hình 11 (a) Phương pháp cơ sở sử dụng HRNet làm xương sống (backbone). (b) Mạng HigherHRNet với giám sát đa độ phân giải (MRS). (c, d) Hệ thống mạng cao hơn với MRS và liên kết tính năng. (e) Mạng HigherHRNet với MRS, tính năng ghép nối và các khối dư thừa.

* Đối với (d) và (e), tập hợp bản đồ nhiệt được sử dụng.



Trong mỗi giai đoạn, HRNet được sử dụng làm xương sống (backbone), có khả năng chính xác hơn và chính xác hơn về mặt không gian so với các mạng có độ phân giải từ cao đến thấp. Để duy trì các biểu diễn độ phân giải cao thông qua ước tính toàn bộ tư thế người đơn lẻ bằng cách kết nối song song các mạng con đa độ phân giải. Bên cạnh đó, việc

hợp nhất nhiều tỷ lệ được lặp lại được thực hiện, điều này làm cho mỗi biểu diễn có độ phân giải từ cao đến thấp nhận được thông tin từ các biểu diễn song song khác. Vì việc ước tính các vị trí trọng điểm của một mạng là khó khăn, nhiều mạng có độ phân giải cao được xếp tầng để tinh chỉnh kết quả đặt ra.

Bản đồ đối tượng được tạo ra bởi mỗi khối có độ phân giải cao được chuyển đến các vị trí tương ứng của giai đoạn tiếp theo bằng cách thực hiện phép cộng. Khác với 4 giai đoạn được sử dụng trong MSPN, ở đây chúng tôi áp dụng mạng 2 giai đoạn do hiệu quả đào tạo và bộ nhớ GPU. Sau, cuối cùng chúng tôi giải mã bản đồ nhiệt (heatmap) và bản đồ bù của giai đoạn cuối để định vị các điểm chính. Heatmap được hồi quy bằng cách sử dụng MSE loss, tương tự như các đường cơ sở đơn giản.

Với $C_k^p = \{x_k^p, y_k^p\}$ là tọa độ của đặc điểm thứ k của đối tượng thứ p cần rút trích. Các đặc điểm của các đối tượng trong cùng khung hình nằm ở các miền phân biệt ngữ nghĩa khác nhau, nên HRNet cần xác định chúng bằng việc che phủ các miền với những Gaussian Kernel có độ lệch chuẩn khác nhau (HRNet tự học và cập nhật độ lệch chuẩn thông qua dữ liệu dùng để training).

Xác định các tỉ lệ s theo hình dạng của heatmap (bản đồ nhiệt). Với mỗi đặc điểm cần rút trích $C_k^p = \{x_k^p, y_k^p\}$, điều chỉnh độ lệch chuẩn thành $\sigma_0 . s_{k,x_k^p,y_k^p}$. Khi đó, vùng được phủ bởi Gaussian Kernel cho đặc điểm cần rút trích trở thành:

$$h_{k,i,j}^p = e^{-\frac{(i-x_k^p)^2+(j-y_k^p)^2}{2(\sigma_0 . s_{k,x_k^p,y_k^p})^2}}$$

Với $\|i - x_k^p\|_1 \leq 3\sigma,\ \|j - y_k^p\|_1 \leq 3\sigma$

Tuy nhiên, vùng phủ này tương đối nhỏ ($s_{k,x_k^p,y_k^p} \sim s_{k,i,j}$) với điều kiện như trên, nên ta có thể thay s_{k,x_k^p,y_k^p} bằng $s_{k,i,j}$

Khi đó ta có bản đồ nhiệt (heatmap) mới, kí hiệu là $H^{\sigma_0 . s}$(heatmap thích nghi tỉ lệ). Mối quan hệ giữa heatmap mới sau khi phủ Gaussian Kernel và heatmap cũ là:

$$H_{k,i,j}^{\sigma_0 . s} = \begin{cases} (H_{k,i,j}^{\sigma_0})^{1/s_{k,i,j}} \ với\ H_{k,i,j}^{\sigma_0} > 0 \\ H_{k,i,j}^{\sigma_0} \ với\ H_{k,i,j}^{\sigma_0} = 0 \end{cases}$$

Với các đặc điểm cần rút trích (ở đây là các khớp đầu, thân,…) có thành phần tỉ lệ lớn hơn 1, độ lệch chuẩn của chúng sẽ vượt σ_0 và phần gaussian kernel phủ sẽ rộng hơn.

Để mô hình ổn định, cần biến đổi một chút để tránh overfitting nhưng vẫn không mất tính tổng quát của mô hình. Hàm mất mát (Loss function) cho các tỉ lệ đã được xác định:

$$L_{reg} = \left\| \left(\frac{1}{s} - 1 \right) 1_{H^{\sigma_0/s} > 0} \right\|_2^2$$

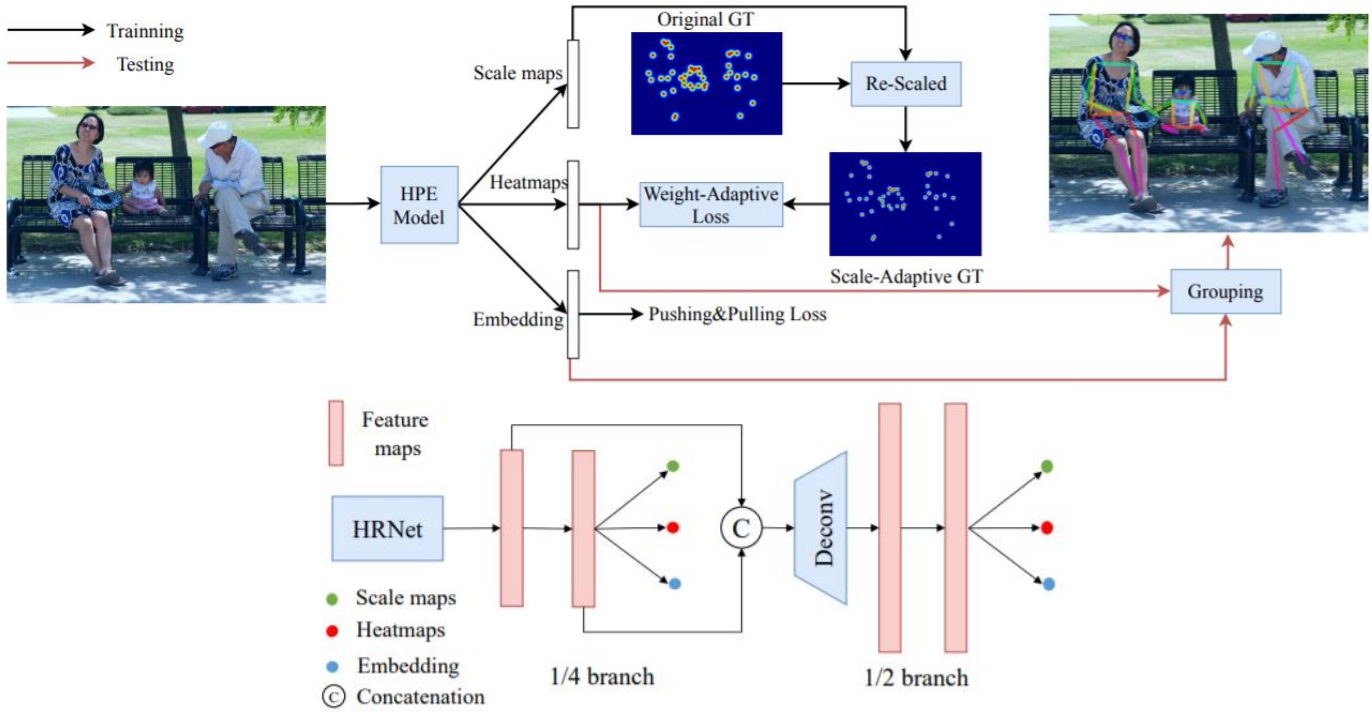
Với $1_{H^{\sigma_0/s}>0}$ chỉ các miền được phủ bởi gaussian kernel. Đồng thời, biến đổi dạng lũy thừa của $H^{\sigma_0 . s}$ thành chuỗi đa thức bằng khai triển Taylor với s = 1 được:

$$H_{k,i,j}^{\sigma_0 . s} = \begin{cases} \frac{1}{2} H_{k,i,j}^{\sigma_0} (1 + (1 + \propto_{k,i,j} ln\left(H_{k,i,j}^{\sigma_0}\right))^2) \ với\ H_{k,i,j}^{\sigma_0} > 0 \\ 0 \ với\ H_{k,i,j}^{\sigma_0} = 0 \end{cases}$$

Với $\propto = \frac{1}{s} - 1$. Khi đó, hàm mất mát tổng là

$$L_{total} = L_{regression} + \lambda L_{reg} = \|P - H^{\sigma_0 . s}\|_2^2 + \lambda \left\| \propto (1_{H^{\frac{\sigma_0}{s}} > 0}) \right\|_2^2$$

(Tham số regularize λ ở đây được chọn là 1 để cân bằng hai đại lượng vế phải)



Hình 12 Giải mã bản đồ nhiệt (heatmap)

Mỗi lần chạy được 1 chủ thể duy nhất. Trong quá trình đào tạo, trước tiên, các bản đồ nhiệt (ground-truth heatmap) được chia tỷ lệ theo bản đồ tỷ lệ dự đoán và sau đó được sử dụng để giám sát toàn bộ mô hình thông qua giảm trọng lượng thích ứng (weight-adaptive loss). Trong quá trình thử nghiệm, các bản đồ nhiệt dự đoán và phương pháp nhúng kết hợp được sử dụng để phân nhóm các cá nhân.

Hình 13 Sau khi tổng hợp (grouping/aggregate) các heatmaps, tags và bouding box.