**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



‘**Logo

Description automatically generated**

**Báo cáo đồ án Môn học**

**CS338 – Nhận Dạng**

**Nhận Vật hoạt hình mô phỏng chuyển động con người**

GVHD:

Sinh viên thực hiện:

Đỗ Văn Tiến

Huỳnh Ngọc Công Danh - 19521322

Phạm Thành Luân - 19520154

Trần Duy Quang - 19522102

Đỗ Minh Trí - 19520307

**TP. HỒ CHÍ MINH – 6/2022**

**Mục lục**

[I. Giới thiệu: 2](#_Toc106659597)

[II. Nội dung: 2](#_Toc106659598)

[1. Giới thiệu về BlazePose: 2](#_Toc106659599)

[2. Giới thiệu về MediaPipe và BlendArMocap: 4](#_Toc106659600)

[3. Phương pháp thực hiện: 5](#_Toc106659601)

[III. Đánh giá: 6](#_Toc106659602)

[1. Ưu điểm của BlazePose: 6](#_Toc106659603)

[2. Nhược điểm: 8](#_Toc106659604)

# Giới thiệu:

Phim hoạt hình là một trong những thể loại phim được nhiều người yêu thích, không phân biệt độ tuổi hay giới tính. Những thước phim vui nhộn, hài hước, nhân vật dễ thương nhưng cũng ẩn chứa nhiều giá trị nhân văn trong đó. Thay vì chỉ là các nét vẽ 2D đơn thuần thì các nhà làm phim đã biết áp dụng các kỹ thuật 3D vào các bộ phim làm cho bộ phim chân thật và bắt mắt hơn.



hình 1. So sánh

Tuy vậy khác với cách làm phim truyền thống việc sản xuất phim hoạt hình là vô cùng vất vả vì ta phải tạo các chuyển động cho các nhân vật không phải ở thế giới thật.

Để dễ dàng tạo các chuyển động các nhà làm phim thường sử dụng các bộ đồ chuyên dụng giúp ghi lại các chuyển động của diễn viên và biển đổi để nó thành chuyển động của các nhân vật ảo. Chi phí của những thiết bị này là khá đắt đỏ nên chúng tôi đề xuất một phương pháp mới với chi phí thấp hơn khi chỉ sử dụng 1 camera bất kì mà không cần dùng đến các bộ đồ chuyên dụng nào.

## 

# Nội dung:

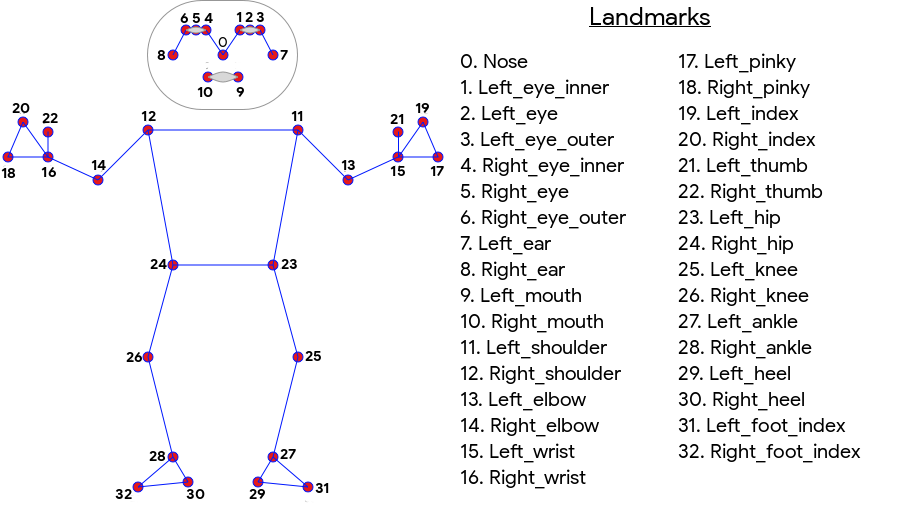
## Giới thiệu về BlazePose:

1. ***Ý tưởng mô hình:***

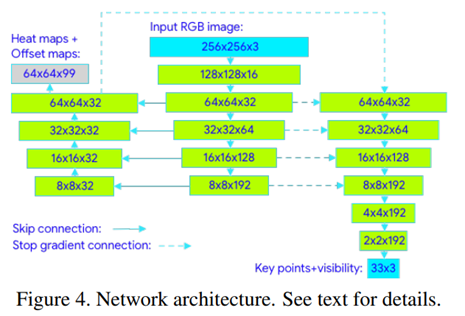
* Mô hình BlazePose gồm 2 phần:
* Bộ Detector: Mô hình gồm một bộ trọng số để phát hiện các điểm (khớp) trên cơ thể trong ảnh
* Bộ Tracker: Mô hình mạng dùng để theo dõi các điểm trên cơ thể
* Sau khi bộ Detector dự đoán ra các điểm (khớp) trên cơ thể người trong ảnh, bộ Tracker sẽ dựa vào frame đó để dự đoán các điểm (khớp) tương ứng ở frame tiếp theo. Nếu như bộ Tracker không cho ra kết quả, bộ Detector sẽ lại dự đoán các điểm (khớp) ở frame đó.

1. ***Kiến trúc mô hình:***

* BlazePose sử dụng cấu trúc liên kết gồm 33 điểm (khớp nối) trên cơ thể bằng cách kết hợp các mô hình:
* BlazeFace: Mô hình dùng để phát hiện khuôn mặt của Google
* BlazePalm: Mô hình dùng để phát hiện khung xương tay của Google
* Microsoft Coco: Mô hình phân loại vật thể sử dụng phương pháp Segmentation

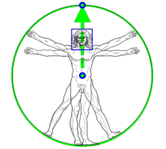


* Về kiến trúc, mô hình gồm 2 phần:
* Encoder-decoder heatmap-based network (dựa trên kiến trúc Stacked Hourglass): Trong đó, Encoder và Decoder được liên kết với nhau bằng Skip-connection (Nhằm cân bằng giữa Low-feature và High-feature) và chỉ sử dụng Decoder trong quá trình training (Loại bỏ trong quá trình testing)
* Regression encoder network: Dự đoán tọa độ của 33 điểm trên không gian 3 chiều và giá trị Visibility (Một điểm có xuất hiện trên frame đó hay không). Trong đó, Heatmap (Output của Decoder heatmap-based network) được sử dụng như một phương pháp học giám sát “yếu” đầu ra của Encoder heatmap-based network . Ngoài ra Encoder heatmap-based network liên kết với Regression encoder network bằng Stop gradient connection (Không cập nhật các tham số này trong training), giúp làm tăng dự đoán của Heatmap và độ chính xác dự đoán của mô hình (Thông qua thực nghiệm cho thấy)

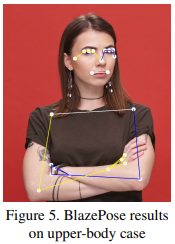


1. ***Quá trình căn chỉnh và tăng cường khớp nối bị che khuất:***

* Trong quá trình training, ta sẽ hạn chế việc biến đổi dữ liệu (Khi thực hiện quá trình Augmentation và Data preparation) như: Hạn chế góc quay ảnh, hạn chế thay đổi kích thước ảnh hay hạn chế biến đổi ảnh. Điều này giúp giảm kích thước mô hình, làm cho mô hình tính toán nhanh hơn và tốn ít tài nguyên để tính toán hơn
* Quá trình Tracking: Dựa vào kết quả của Detector hay kết quả frame trước đó, ta sẽ căn chỉnh cơ thể người để điểm chính giữa hông ngay trung tâm của ảnh vuông (Input network: 256x256x3 Image). Sau đó ta sẽ xoay đường nối giữa điểm hông và điểm chính giữa 2 điểm vai sao cho đường đó song song với trục y của frame trước đó. Cuối cùng tỉ lệ được ước tính sao cho tất cả các điểm trên cơ thể giới hạn trong 1 hình vuông hộp (Không gian 3 chiều) bao quanh cơ thể

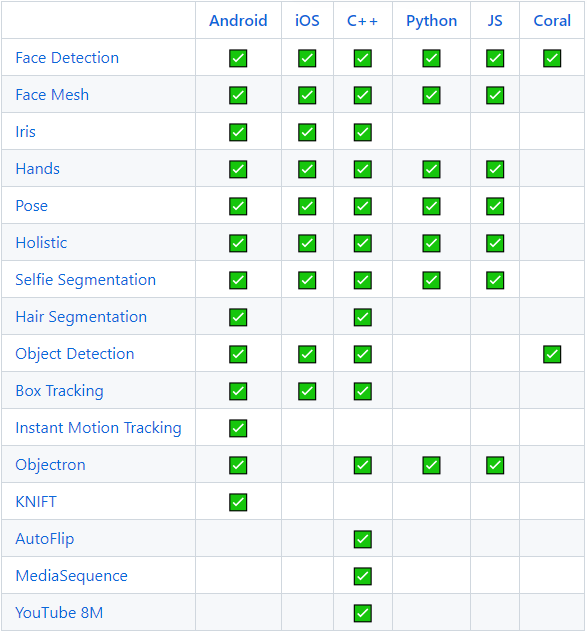


* Để đưa ra các điểm khớp nối tại các bộ phận bị che khuất hoặc không xuất hiện trong frame ảnh, ta sử dụng phương pháp tự giả lập ra các điểm khớp nối đó. Trong quá trình training, sẽ có một phương pháp (Per-point visibility classifier), phương pháp này sẽ phân loại 1 khớp là occluded (Bị che khuất) hay dự đoán sai (*Có thể hiểu, trong quá trình training, nếu xuất hiện các khớp nối bị che khuất thì mô hình sẽ bỏ qua và không đánh giá độ sai số, còn khi 1 khớp bất kì xuất hiện, lúc đó chương trình mới đánh giá độ chính xác của dự đoán của mô hình đó đối với khớp nối đó. Còn trong quá trình test, đơn giản là mô hình sẽ dự đoán khớp nối của các phần cơ thể xuất hiện trong frame ảnh, còn thiếu khớp nào thì mô hình tự simulate ra)*



## Giới thiệu về MediaPipe và BlendArMocap:

* MediaPipe là tập hợp của một loạt các giải pháp Machine Learning đa nền tảng, được phát triển bởi Google. MediaPipe được sử dụng để giải quyết nhiều bài toán như: Face detection, Hand detection, Pose estimation, Object tracking,...



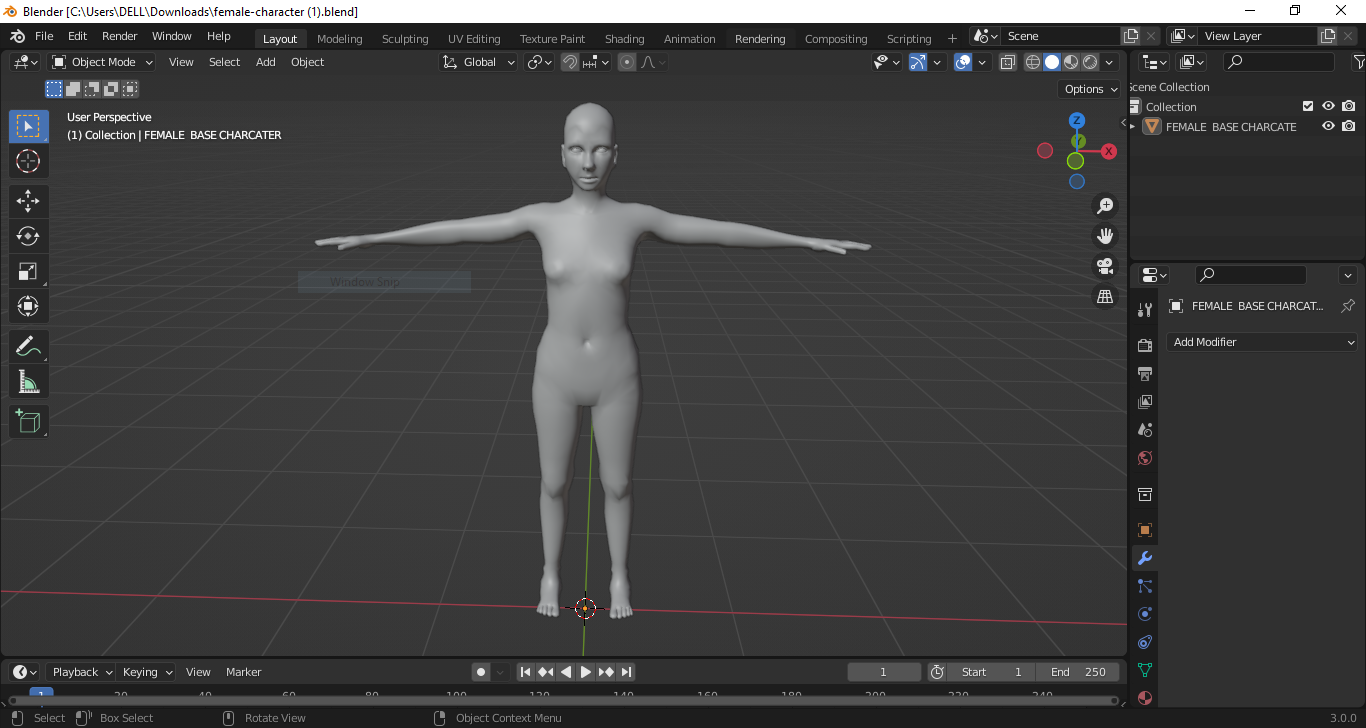
* BlendArMocap là một package (Hay Blender add-on), được dùng để phát hiện khuôn mặt, khung xương tay hay cơ thể chỉ với sử dụng Webcam. Các mô hình của BlendArMocap đều sử dụng dựa trên các mô hình có sẵn của MediaPipe. Do đó, với Pose estimation, BlendArMocap sử dụng mô hình BlazePose với cấu trúc 33 điểm (khớp nối) để phát hiện khung xương cơ thể. Sau đó, thông tin phát hiện này sẽ được dùng làm Rigify trong Blender, từ đó có thể điều chỉnh, thay đổi các khớp nối trong mô hình.



## 

## Phương pháp thực hiện:

* Quy trình thực hiện:
* Sử dụng Blender tạo mô hình khung xương và mô hình Pose estimation của MediaPipe để detect, tracking cơ thể người có trong video

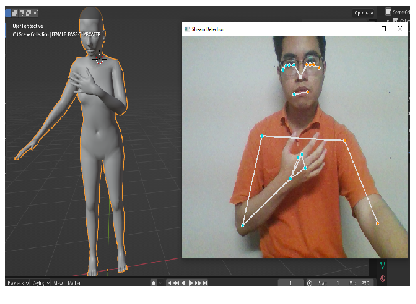


* Khi sử dụng MediaPipe cho bài toán Pose estimation, ta sẽ sử dụng mô hình gồm 33 điểm trên cơ thể được sử dụng trong BlazePose. Mỗi điểm bao gồm 4 giá trị gồm: 3 giá trị biểu thị cho tọa độ của điểm đó trong không gian 3 chiều (Với miền giá trị trong khoảng (0, 1)) và giá trị còn lại biểu thị mô hình Detector có phát hiện ra điểm đó trên cơ thể hay không (Visibility). Sau đó, đọc các tọa độ này vào mô hình trên Blender, điều chỉnh hệ trục tọa độ của mô hình Pose estimation sao cho khớp với hệ trục tọa độ mô hình Blender

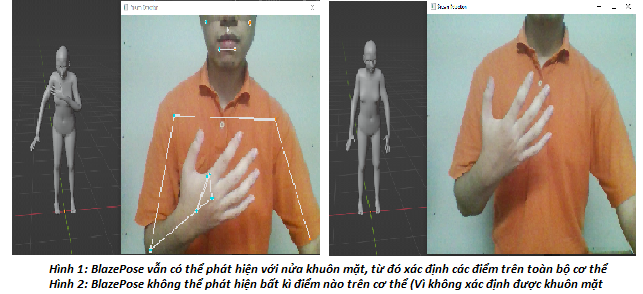
# Đánh giá:

## Ưu điểm của BlazePose:

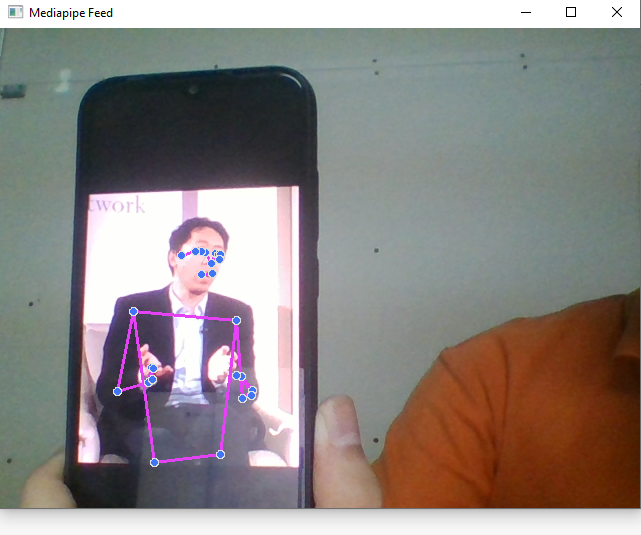
* Mô hình BlazePose có thể phát hiện được cơ thể người trong ảnh nhanh chóng. Nguyên nhân vì:
* Trong bộ Detector, mặc dù mô hình mạng gồm Encoder-decoder heatmap-based network, nhưng phần Decoder trong đó chỉ sử dụng trong quá trình training và bị lược bỏ khi thực hiện testing hay Feedforward, do đó mô hình có thể tận dụng tối đa công năng của Heatmap mà không phải thêm quá nhiều trọng số.
* Ngoài ra, trong Training set, các phép biến đổi ảnh không quá nhiều nên điều này cũng giảm sự phức tạp của mô hình.
* Cuối cùng, quá trình Simulate điểm (khớp nối) bị che (hay không xuất hiện) trong mô hình khá đơn giản, cho thấy mô hình không quá quan trọng việc tính toán các điểm đó. Đây chính là nguyên nhân khiến cho các mô hình trước đây không thể xử lý trong real time (Vì các mô hình trước đó quá tập trung vào việc tính toán ra vị trí chính xác của các điểm bị khuất)



* Mô hình có thể giải quyết các trường hợp một phần cơ thể bị che mất (Ngoại trừ khuôn mặt). Điều này đã được nêu ở trên. Tuy nhiên, với các frame không có khuôn mặt người, mô hình sẽ không detect được toàn bộ cơ thể. Nguyên nhân vì BlazePose sẽ dùng Face detector (Vì tác giả cho rằng khuôn mặt người là bộ phận ít bị biến đổi hình dạng nhất, khác với tay hoặc chân), sau đó mới xác định toàn bộ cơ thể. Ngoài ra, bộ Dataset của tác giả chỉ gồm các frame ảnh mà có xuất hiện toàn bộ cơ thể người hoặc ít nhất phải xuất hiện khuôn mặt (Có nêu trong paper)

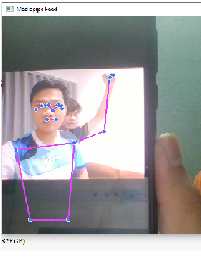


* Kích thước mô hình nhỏ (đã giải thích trên), do đó mô hình có thể chạy real-time và có thể chạy tốt trên các thiết bị không có nhiều tài nguyên tính toán (Vd: Điện thoại, Camera,... )
* Mô hình vẫn có thể cho ra kết quả phát hiện khung xương ở các trường hợp không phải vật thể sống (Tức một bức ảnh người trong frame ảnh)



## Nhược điểm:

* Mô hình BlazePose chỉ hoạt động tốt khi trong ảnh chỉ có một người. Nguyên nhân vì: Trong quá trình tracking, ở frame trước đó, cơ thể người sẽ được căn chỉnh sao cho điểm giữa hông nằm ở vị trí trung tâm ảnh. Do đó mô hình không thể hoạt động hoặc hoạt động bị sai khi ảnh có nhiều người.

****

* Mô hình vẫn còn sử dụng quá nhiều phương pháp biến đổi xử lý ảnh. Cụ thể như các phép căn chỉnh ảnh sao cho điểm giữa hông nằm ở trung tâm ảnh đầu vào. Hay các phép ước tính vị trí các điểm bị che khuất vẫn còn dựa trên công thức cố định. Điều này mặc dù giúp cho mô hình trở nên nhẹ hơn và chạy nhanh hơn, tính toán ít hơn, nhưng cũng đánh đổi đi độ chính xác thực tế trên dữ liệu