TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO TIẾN ĐỘ GIỮA KỲ

# SINH TRẮC HỌC

***CHỦ ĐỀ: NHẬN DIỆN DÁNG ĐI***

## 1 THÔNG TIN CHUNG

**Giảng viên hướng dẫn:**

- PGS. TS. Lê Hoàng Thái (Khoa Công nghệ thông tin)
- ThS Dương Thái Bảo (Khoa Công nghệ thông tin)

**Nhóm sinh viên thực hiện là nhóm 7, gồm:**

1. Phạm Thái Huy (MSSV: 21120081)
2. Nguyễn Đức Mạnh (MSSV: 22120204)
3. Lê Quang Vĩnh Quyền (MSSV: 22120307)

# Mục lục

# 2 NỘI DUNG BÁO CÁO

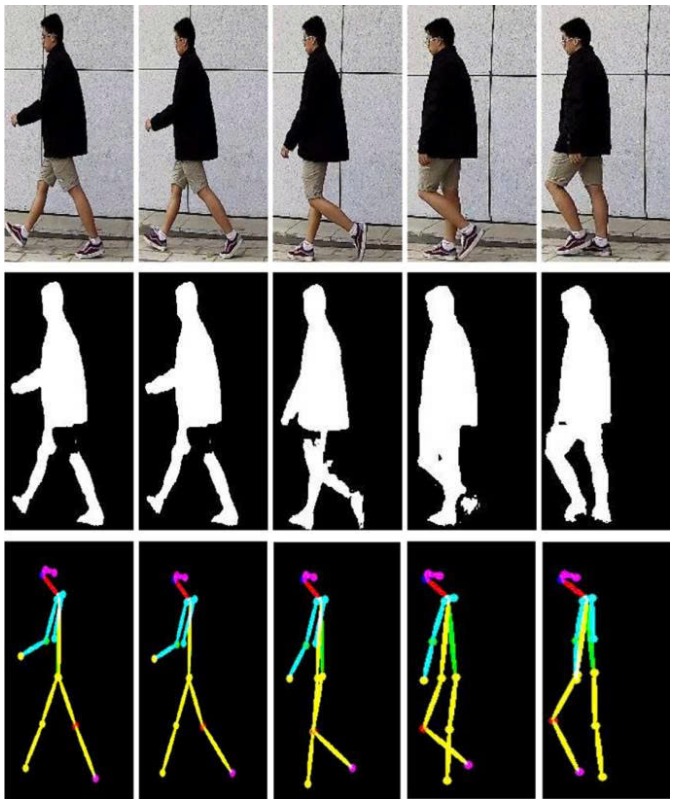
## 2.1 Tổng hợp nội dung chương sách được chọn

### 2.1.1 Giới thiệu

Với mục tiêu định danh con người từ khoảng cách xa và trong môi trường không bị ràng buộc, nhận diện dáng đi tập trung phân tích và xác định danh tính của cá nhân dựa trên các kiểu đi bộ đặc trưng hoặc cách thức vận động của họ qua các chuỗi hình ảnh được thu thập từ camera. Khác với các đặc điểm sinh trắc học vật lý truyền thống như dấu vân tay, khuôn mặt hay mống mắt, dáng đi được xếp vào nhóm sinh trắc học hành vi. Đặc điểm này được hình thành từ cấu trúc cơ xương độc nhất và thói quen vận động của mỗi người, tạo nên một dấu ấn riêng biệt có khả năng phân biệt cao. Ngay từ những năm 1970, các nghiên cứu tâm lý học nền tảng của Cutting và Kozlowski đã chứng minh rằng con người có khả năng nhận diện người quen chỉ thông qua sự chuyển động của các điểm sáng mô phỏng dáng đi mà không cần bất kỳ thông tin chi tiết nào về ngoại hình hay khuôn mặt.

Một trong những lợi thế quan trọng nhất giúp nhận diện dáng đi trở nên nổi bật là khả năng hoạt động hiệu quả trong các điều kiện khó khăn. Trong khi nhận diện khuôn mặt hay mống mắt yêu cầu đối tượng phải hợp tác, đứng gần thiết bị thu nhận và cần hình ảnh độ phân giải cao, nhận diện dáng đi có thể thực hiện từ khoảng cách xa, có thể lên tới hàng trăm mét và chấp nhận dữ liệu có độ phân giải thấp. Đặc biệt, đây là một phương thức nhận diện không xâm lấn, nghĩa là hệ thống có thể thu thập dữ liệu thụ động qua camera giám sát mà không cần sự tương tác trực tiếp hay sự chủ động hợp tác của đối tượng. Hơn nữa, do dáng đi là một hành vi vô thức bắt nguồn từ cơ chế sinh học, do đó việc ngụy trang hay giả mạo dáng đi trong thời gian dài là vô cùng khó khăn đối với các đối tượng muốn che giấu danh tính.

Nhìn chung, việc ứng dụng hệ thống nhận diện dáng đi hiện thực phải đối mặt



Hình 1: Các loại dữ liệu cho bài toán nhận diện dáng đi.

với rất nhiều thử thách lớn đến từ các yếu tố ngoại cảnh. Độ chính xác của hệ thống thường sẽ rất dễ bị ảnh hưởng nghiêm trọng bởi sự thay đổi của góc nhìn của camera, đây được xem là yếu tố gây nhiễu lớn nhất làm thay đổi hình dạng hình học của đối tượng trên khung hình. Bên cạnh đó, các điều kiện về trang phục như áo khoác dày che khuất cơ thể, hoặc trạng thái mang vác vật dụng như ba lô, túi xách cũng làm thay đổi trọng tâm và biên độ dao động của dáng đi. Các yếu tố môi trường khác như bề mặt đường đi, điều kiện ánh sáng hay sự che khuất bởi vật cản cũng góp phần làm giảm hiệu suất nhận diện khi áp dụng vào các tình huống đời sống.

### 2.1.2 Vấn đề

Bài toán thử thách HumanID được thiết lập trong chương trình HumanID At a Distance của DARPA nhằm cung cấp một khung tham chiếu khách quan và định lượng để đo lường tiến bộ trong nghiên cứu nhận dạng dáng đi. Bài toán bao gồm

ba thành phần chính: cơ sở dữ liệu quy mô lớn, hệ thống các thí nghiệm thử thách với độ khó tăng dần và một thuật toán cơ sở

**a. Cơ sở dữ liệu và biến số** Dữ liệu bao gồm 1870 chuỗi video từ 122 cá nhân, được thu thập trong môi trường ngoài trời. Mỗi đối tượng được ghi hình dưới sự kết hợp của năm biến số chính:

- **Góc quay camera:** Hai góc nhìn bên trái và bên phải.
- **Loại giày:** Hai loại giày khác nhau.
- **Bề mặt:** Cỏ và bê tông.
- **Vật dụng mang theo:** Có mang túi xách hoặc không mang.
- **Thời gian:** Hai thời điểm thu thập cách nhau 6 tháng.

Ngoài ra, bộ dữ liệu còn cung cấp các thông tin bổ trợ như nhân trắc học (tuổi, giới tính, chiều cao) và các hình ảnh bóng được tách thủ công cho 71 đối tượng để hỗ trợ nghiên cứu lỗi phân đoạn và xây dựng mô hình chi tiết các bộ phận cơ thể.

**b. Baseline Gait Algorithm** Thuật toán cơ sở dựa trên việc khớp mẫu hình bóng. Quá trình này được mô tả thông qua các bước toán học sau:

1. **Ước lượng chu kỳ:** Xác định chu kỳ dáng đi ($N_{Gait}$) dựa trên sự biến thiên định kỳ của số lượng pixel tiền cảnh ở phần dưới hình bóng. Gọi $f(t)$ là hàm đếm số pixel tại khung hình $t$, chu kỳ được tính dựa trên khoảng cách giữa các điểm cực tiểu liên tiếp của $f(t)$.
2. **Độ đo tương quan không gian - thời gian:** Giả sử chuỗi kiểm tra là $S_P = \{S_P(1), \ldots, S_P(M)\}$ và chuỗi chuẩn (Gallery) là $S_G = \{S_G(1), \ldots, S_G(N)\}$. Quá trình so sánh thực hiện qua các bước:
   - *Độ tương đồng khung hình:* Sử dụng chỉ số Tanimoto ($S$) để so sánh hai

khung hình $i$ và $j$:

$$S(S_P(i), S_G(j)) = \frac{\#(S_P(i) \cap S_G(j))}{\#(S_P(i) \cup S_G(j))} \tag{1}$$

Trong đó, # là số lượng pixel tiền cảnh của phép giao ($\cap$) và phép hợp ($\cup$).

- *Tương quan chuỗi con:* Chia $S_P$ thành các chuỗi con $S_{Pk}$ có độ dài $N_{Gait}$. Độ tương quan giữa $S_{Pk}$ và $S_G$ tại vị trí dịch chuyển $l$ là:

$$\text{Corr}(S_{Pk}, S_G)(l) = \sum_{j=1}^{N_{Gait}} S(S_P(k+j), S_G(l+j)) \tag{2}$$

3. **Hàm độ đo tương đồng cuối cùng:** Để chống nhiễu và sai số phân đoạn cục bộ, độ tương đồng tổng thể ($Sim$) được tính bằng giá trị trung vị của các giá trị tương quan cực đại:

$$Sim(S_P, S_G) = \text{Median}_k \left( \max_l \text{Corr}(S_{Pk}, S_G)(l) \right) \tag{3}$$

**c. Các thách thức trong thực nghiệm (Challenge Experiments)** Để đánh giá hiệu suất của thuật toán dưới tác động của các biến số khác nhau, bài toán thiết lập 12 thách thức trong thí nghiệm. Một tập hợp chuẩn (*Gallery*) được cố định làm nhóm đối chứng với các điều kiện: bề mặt cỏ, giày loại A, góc nhìn bên phải, không mang túi xách và dữ liệu thu thập vào tháng 5.

Các tập kiểm tra được thiết kế để cô lập hoặc kết hợp các biến số nhằm đo lường sự sụt giảm hiệu suất:

- **Các thí nghiệm đơn biến :** Được đánh dấu () trong bảng thử thách, bao gồm:
  - **Exp A (View):** Thay đổi góc nhìn (Trái vs Phải).
  - **Exp B* (Shoe):** Thay đổi loại giày (A vs B).
  - **Exp D* (Surface):** Thay đổi bề mặt đi bộ (Cỏ vs Bê tông).

  - **Exp H* (Carry):** Thay đổi điều kiện mang vác (Có túi xách vs Không).
  - **Exp K* (Time):** Thay đổi thời gian thu thập (Tháng 5 vs Tháng 11), bao gồm cả thay đổi ngầm định về trang phục.

- **Các thí nghiệm đa biến:** Các thí nghiệm còn lại (như C, E, F, G, I, J, L) kết hợp từ hai đến ba biến số cùng lúc như là vừa thay đổi bề mặt, vừa thay đổi góc nhìn để kiểm tra tính bền vững của hệ thống trong các tình huống phức tạp hơn.

Việc phân cấp các thí nghiệm này cho phép xác định thứ tự độ khó của bài toán nhận dạng. Thông thường, các thí nghiệm liên quan đến bề mặt và thời gian được coi là những thử thách lớn nhất đối với các thuật toán nhận dạng dáng đi hiện nay.

**d. Đánh giá hiệu suất** Kết quả thực nghiệm trên bài toán thử thách HumanID phản ánh sự tiến bộ vượt bậc của các thuật toán nhận dạng dáng đi qua hai giai đoạn phát triển chính:

- **Giai đoạn khởi đầu (2002):** Khi bài toán mới được công bố, thuật toán cơ sở (Baseline) đạt hiệu suất cao hơn hầu hết các thuật toán thế hệ đầu. Điều này cho thấy tính hiệu quả của phương pháp khớp mẫu hình bóng đơn giản nhưng ổn định.
- **Giai đoạn cải tiến (2004 – 2006):** Hiệu suất của các thuật toán mới bắt đầu vượt xa Baseline một cách đáng kể. Sự cải thiện này không chỉ đến từ việc tối ưu hóa kỹ thuật (engineering) mà còn nhờ sự thay đổi trong tư duy tiếp cận: chuyển từ phân tích động lực học thuần túy sang phân tích hình dạng hình bóng (silhouette shapes) và các mô hình dựa trên bộ phận cơ thể.

Dựa trên các kết quả báo cáo, có thể rút ra các nhận định quan trọng sau:

- **Độ khó của biến số:** Các thí nghiệm thay đổi bề mặt (Exp D) và thời gian (Exp K) vẫn là những thách thức lớn nhất. Mặc dù hiệu suất đã được cải

thiện, nhưng khoảng cách giữa kết quả thực tế và kỳ vọng vẫn còn lớn so với các thí nghiệm về góc nhìn hay loại giày.

- **Tính đa dạng của phương pháp:** Các nghiên cứu cho thấy động lực học là quan trọng nhưng chưa đủ; việc kết hợp thông tin về hình thái (morphology) giúp hệ thống bền vững hơn trước các sai số phân đoạn và nhiễu hậu cảnh.

#### 2.1.3 Phương pháp

Hầu hết các phương pháp nhận dạng dáng đi đều dựa trên việc khai thác hình bóng (silhouette) của đối tượng. Đây là đặc trưng mức thấp phổ biến nhờ tính bền vững trước sự thay đổi màu sắc trang phục và khả năng trích xuất dễ dàng trong môi trường camera tĩnh. Các hướng tiếp cận chính được chia thành ba nhóm:

**a. Phương pháp dựa trên căn chỉnh thời gian** Đây là hướng tiếp cận phổ biến nhất, coi chuỗi dáng đi là một chuỗi thời gian chứa đựng cả đặc trưng hình dạng và động lực học. Quá trình này thường bao gồm ba giai đoạn: trích xuất đặc trưng (silhouette, PCA, mô tả Fourier), căn chỉnh chuỗi và tính toán độ đo.

- **Căn chỉnh chuỗi:** Sử dụng các kỹ thuật như Biến đổi thời gian động (Dynamic Time Warping - DTW) để đồng bộ hóa các chuỗi có tốc độ khác nhau, hoặc Mô hình Markov ẩn (HMM) để mô tả sự chuyển giao giữa các tư thế.
- **Mô hình toán học HMM:**

$$P(O|\lambda) = \sum_{q} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} \dots b_{q_T}(o_T) \tag{4}$$

**Trong đó:**

  - $O = \{o_1, o_2, \dots, o_T\}$ là chuỗi các đặc trưng hình bóng quan sát được từ video.
  - $\pi_{q_1}$: Xác suất bắt đầu tại trạng thái tư thế $q_1$.
  - $a_{q_{t-1}q_t}$: Xác suất chuyển đổi giữa hai tư thế liên tiếp (ví dụ: từ tư thế chân chụm sang chân xòe). Nó mô tả **động lực học** của bước đi.

– $b_{q_t}(o_t)$: Xác suất để một hình bóng cụ thể $o_t$ xuất hiện tại trạng thái $q_t$. Nó mô tả **đặc điểm hình dạng** của tư thế đó.

– $P(O|\lambda)$ là tổng xác suất của chuỗi video khớp với mô hình mẫu $\lambda$ của một cá nhân.

**b. Phương pháp dựa trên hình dạng** Hướng tiếp cận này ưu tiên sự tương đồng về hình thái của hình bóng và giảm nhẹ yếu tố trình tự thời gian. Một số kỹ thuật tiêu biểu bao gồm:

- **Hình ảnh năng lượng dáng đi (Gait Energy Image - GEI):**

$$A(x, y) = \frac{1}{N} \sum_{t=1}^{N} S(x, y, t) \tag{5}$$

**Trong đó:**

– $S(x, y, t)$ là giá trị pixel (0 hoặc 1) tại tọa độ $(x, y)$ của khung hình thời điểm $t$.

– $A(x, y)$ (GEI) là một ảnh xám duy nhất đại diện cho toàn bộ chu kỳ.

– Các vùng có cường độ sáng cao nhất trong GEI cho thấy đó là nơi cơ thể ít thay đổi nhất (thường là đầu và thân), trong khi các vùng mờ (cường độ thấp) đại diện cho các bộ phận chuyển động mạnh như tay và chân.

- **Chuẩn hóa động lực học :** Sử dụng một mô hình HMM chung cho toàn bộ quần thể để ánh xạ các khung hình vào các khung hình tư thế chuẩn (*stance-frames*). Khoảng cách giữa hai đối tượng sau đó được tính toán trong không gian.

- **Phân tích biệt thức tuyến tính (LDA):**

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{6}$$

**Trong đó:**

– $S_B$: Đo lường sự khác biệt giữa các cá nhân khác nhau. Chúng ta muốn giá trị này lớn nhất để dễ phân biệt người này với người kia.
– $S_W$: Đo lường sự biến thiên của cùng một người dưới các điều kiện khác nhau (ví dụ: cùng một người nhưng khi mặc áo khoác hoặc đi trên cỏ).

**c. Phương pháp tham số tĩnh** Hướng tiếp cận này trích xuất các thông số hình học trực tiếp từ cơ thể và các đặc trưng vận động cơ bản:

– **Tham số vận động:** Độ dài sải chân , tốc độ bước đi và nhịp độ .
– **Tham số hình thể:** Tỷ lệ kích thước các bộ phận cơ thể (tỷ lệ chiều dài chân/thân).

Tuy nhiên, phương pháp này thường đòi hỏi việc hiệu chuẩn camera 3D phức tạp và có hiệu suất thấp hơn khi xử lý dữ liệu ở khoảng cách xa hoặc độ phân giải thấp.

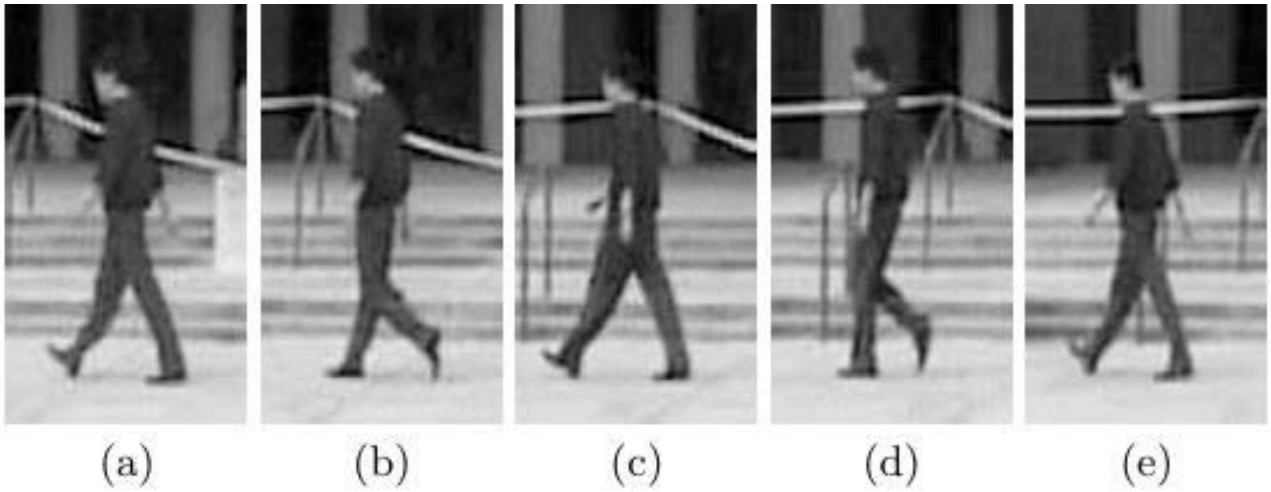
#### 2.1.4 Thảo luận và Hướng nghiên cứu tiếp theo

**Hình dáng và động lực học của dáng đi**

Trong nghiên cứu về dáng đi, sự tranh luận giữa vai trò của hình dáng và động lực học luôn là tiêu điểm. Các nghiên cứu thực nghiệm ban đầu chỉ ra rằng con người có khả năng nhận diện danh tính dựa trên các đặc điểm động lực học ngay cả khi hình dáng bị che khuất, tuy nhiên các phương pháp phân tích dựa trên hình dáng hình bóng lại mang lại hiệu quả vượt trội. Nhiều thuật toán tiên tiến gần đây đã chứng minh rằng việc tập trung vào hình dáng của từng giai đoạn trong chu kỳ bước đi giúp hệ thống đạt được độ chính xác cao hơn, đặc biệt là khi phải đối mặt với các biến số khó như thay đổi bề mặt đi bộ. Mặc dù vậy, động lực học vẫn đóng vai trò không thể thiếu vì nó chứa đựng các thông tin về tốc độ và sự chuyển tiếp giữa các pha vận động vốn mang tính đặc trưng cho từng cá nhân. Tuy nhiên, các nghiên cứu mới gần đây chỉ ra rằng việc phụ thuộc quá nhiều vào hình dáng sẽ khiến hệ thống dễ bị sai lệch khi đối tượng thay đổi trang phục. Do

đó, xu hướng hiện tại là phát triển các mô hình học biểu diễn không gian - thời gian phân cấp, cho phép tách biệt các đặc trưng chuyển động cốt lõi ra khỏi các đặc điểm hình dáng bề ngoài dễ thay đổi, từ đó tận dụng sức mạnh của cả hai yếu tố này.

Hình 2 minh họa chu kỳ dáng đi được chia thành các giai đoạn khác nhau, thể hiện rõ sự chuyển động tuần hoàn của các bộ phận cơ thể trong quá trình đi bộ.




Hình 2: Minh họa chu kỳ dáng đi được chia thành bốn giai đoạn: (i) tựa chân phải; (ii) lăng chân trái; (iii) tựa chân trái; và (iv) lăng chân phải, tương ứng với các trạng thái từ (a) đến (e). Khoảng thời gian cả hai chân cùng tiếp xúc mặt sàn được gọi là giai đoạn hỗ trợ kép.

**Chất lượng hình bóng và nhận dạng dáng đi**

Chất lượng của các hình bóng phụ thuộc vào khả năng phân biệt giữa nền và đối đượng. Trong môi trường ngoài trời, các yếu tố nhiễu như bóng đổ, thay đổi ánh sáng và sự chuyển động của hậu cảnh khiến việc tách hình bóng trở nên cực kỳ khó khăn. Tuy nhiên, một phát hiện chỉ ra rằng sự sụt giảm hiệu suất khi thay đổi bề mặt hoặc thời gian không hoàn toàn do lỗi xử lý hình bóng ở cấp độ thấp gây ra. Ngay cả khi sử dụng các hình bóng đã được tiền xử lý thủ công, kết quả vẫn cho thấy sự suy giảm đáng kể, nghĩa là bản thân dáng đi của con người đã có những thay đổi cơ bản khi điều kiện môi trường thay đổi. Điều này cho thấy các công trình tương tai thay vì tìm kiếm các phương pháp tốt hơn để phát hiện hình bóng nhằm cải thiện nhận dạng, việc nghiên cứu và tách biệt các thành phần của

dáng đi không thay đổi theo giày dép, bề mặt hoặc thời gian sẽ hiệu quả hơn.

**Các biến số**

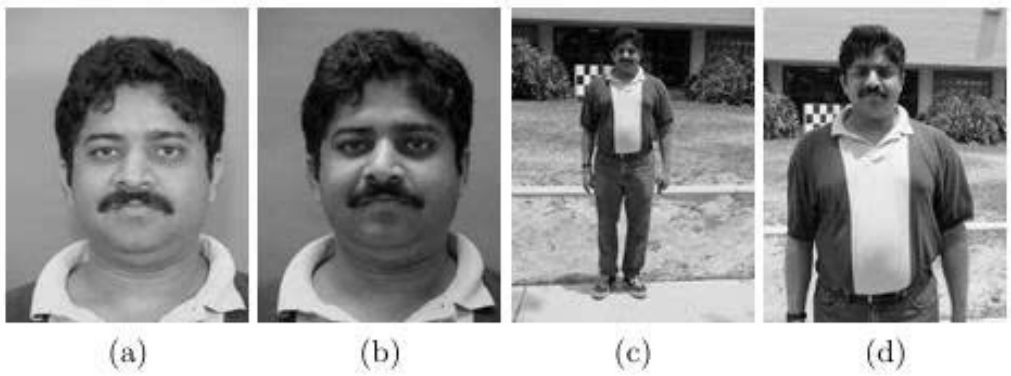
Thách thức của nhận diện dáng đi nằm ở việc duy trì độ ổn định trước các tác động của các yếu tố ngoại cảnh. Trong khi các yếu tố như loại giày dép hay việc mang theo túi xách có tác động tương đối nhỏ, thì sự thay đổi về bề mặt đi bộ và khoảng cách thời gian giữa các lần thu thập dữ liệu lại gây ra những ảnh hưởng tiêu cực nghiêm trọng. Đặc biệt, nhận dạng dáng đi theo thời gian là một bài toán khó do sự thay đổi về trang phục theo mùa và những biến đổi tự nhiên trong cơ thể người theo thời gian. Ngoài ra, việc di chuyển trong các môi trường thực tế với các góc nhìn camera đa dạng, vật cản che khuất và độ phân giải thấp vẫn là những trở ngại cần được giải quyết. Do đó các nghiên cứu trong tương lai cần tập trung vào việc mô hình hóa các thành phần dáng đi không thay đổi hoặc tìm cách dự đoán sự biến đổi của dáng đi khi chuyển từ bề mặt này sang bề mặt khác.

**Các bộ dữ liệu trong tương lai**

Sự phát triển của nhận dạng dáng đi phụ thuộc rất lớn vào quy mô và độ đa dạng của các bộ dữ liệu. Các chuyên gia nhận định rằng cần có những bộ dữ liệu khổng lồ với quy mô lên tới hàng nghìn đối tượng để hỗ trợ việc hiểu sâu hơn về sự biến thiên của dáng đi trong điều kiện ngoài trời và theo thời gian. Các bộ dữ liệu chuẩn như CASIA-B hay OU-MVLP đang dần trở nên bão hòa khi các mô hình Học sâu đã đạt độ chính xác rất cao trên đó. Hiện nay, các bộ dữ liệu như GREW hay Gait3D đã bắt đầu chuyển hướng từ môi trường phòng thí nghiệm sang môi trường thực tế với hàng chục nghìn danh tính và hàng triệu chuỗi hình ảnh. Một hướng đi mới đầy triển vọng là sử dụng dữ liệu tổng hợp được tạo ra từ các mô hình cơ thể người ảo, giúp giải quyết vấn đề thiếu hụt dữ liệu được dán nhãn và giảm bớt các rào cản về chi phí thu thập dữ liệu thực tế. Đồng thời, việc khai thác dữ liệu video không nhãn trên quy mô lớn thông qua các phương pháp học tự giám sát đang trở thành một lĩnh vực nghiên cứu đầy tiềm năng.

**Kết hợp khuôn mặt và dáng đi**

Mặc dù dáng đi có ưu thế ở khoảng cách xa, thì việc kết hợp giữa nhận dạng dáng đi và khuôn mặt là một giải pháp tối ưu để nâng cao độ tin cậy của hệ thống. Dáng đi có thể được sử dụng khi đối tượng ở khoảng cách xa, trong khi khuôn mặt sẽ phát huy tác dụng khi đối tượng tiến lại gần camera hơn. Việc kết hợp đa phương thức dáng đi và mặt người mang lại hiệu suất vượt trội so với việc chỉ sử dụng một loại sinh trắc học đơn lẻ, đồng thời giúp hệ thống bền bỉ hơn trước các nhiễu động của từng loại dữ liệu. Hình 3 minh họa các mẫu khuôn mặt dưới nhiều điều kiện khác nhau, cho thấy sự đa dạng trong dữ liệu sinh trắc học. Trong tương lai, việc tích hợp này không chỉ dừng lại ở khuôn mặt mà còn có thể mở rộng sang các phương thức khác như thông tin về chiều cao, kích thước các bộ phận cơ thể hoặc dữ liệu từ nhiều góc nhìn camera cùng lúc để tạo ra một hồ sơ định danh toàn diện và chính xác hơn.




Hình 3: Các mẫu khuôn mặt dưới nhiều điều kiện khác nhau: (a) và (b) là ảnh tập mẫu trong điều kiện ánh sáng khác nhau; (c) và (d) là ảnh tập kiểm tra chụp ngoài trời ở khoảng cách xa và gần.

### Quyền riêng tư và bảo mật sinh trắc học

Một khía cạnh mới đang ngày càng được chú trọng là tính bảo mật và quyền riêng tư của dữ liệu dáng đi. Do dáng đi có thể được thu thập từ xa mà không cần sự hợp tác hay nhận biết của đối tượng, công nghệ này làm dấy lên những lo ngại nghiêm trọng về quyền riêng tư cá nhân và sự giám sát đại chúng. Bên cạnh đó, vấn đề an ninh của chính hệ thống AI cũng đáng báo động với sự xuất hiện của các cuộc tấn công đối kháng, nơi kẻ tấn công có thể thay đổi một chút dáng đi

hoặc thêm nhiễu vào video để đánh lừa hệ thống. Các luật lệ như quy định chung về Bảo vệ Dữ liệu Châu Âu (GDPR) đã đặt ra những hạn chế chặt chẽ đối với việc sử dụng dữ liệu sinh trắc học, thúc đẩy cộng đồng nghiên cứu phải tìm kiếm các giải pháp bảo vệ quyền riêng tư. Các hướng nghiên cứu mới bao gồm việc phát triển các phương pháp ẩn danh hóa, mã hóa video dáng đi sao cho hệ thống nhận dạng vẫn hoạt động nhưng danh tính con người không thể bị quan sát bằng mắt thường, cũng như tăng cường khả năng chống lại các cuộc tấn công giả mạo hoặc tấn công đối nghịch nhằm đánh lừa hệ thống nhận dạng.

### 2.1.5 Kết luận

Nhìn lại toàn bộ quá trình phát triển, nhận dạng dáng đi đã khẳng định vị thế là một trong những công nghệ sinh trắc học tiềm năng nhất, với điểm mạnh về khả năng định danh tầm xa và không xâm lấn mà các phương pháp truyền thống như khuôn mặt hay vân tay không thể thay thế. Sự chuyển dịch mạnh mẽ từ các kỹ thuật thị giác máy tính cổ điển sang các kiến trúc Học sâu tiên tiến đã nâng tầm lĩnh vực này, giúp các hệ thống hiện đại đạt được độ chính xác ấn tượng trên các bộ dữ liệu tiêu chuẩn. Các nghiên cứu đột phá gần đây đã chứng minh rằng việc khai thác sâu các biểu diễn không gian - thời gian phân cấp là chìa khóa để tách biệt các đặc trưng vận động cốt lõi khỏi những yếu tố nhiễu loạn của bề mặt, mở ra triển vọng to lớn trong việc giải mã hành vi vận động của con người.

Tuy nhiên, thực tế triển khai đã chỉ ra một khoảng cách đáng kể về tính thực tiễn giữa môi trường phòng thí nghiệm lý tưởng và thế giới thực hỗn loạn. Như một vài phân tích thực nghiệm đã làm rõ hiệu suất của các thuật toán hàng đầu vẫn sụt giảm nghiêm trọng khi đối mặt với sự đa dạng không giới hạn của các biến số ngoại cảnh như góc quay camera an ninh phức tạp, sự thay đổi trang phục theo mùa, hay điều kiện ánh sáng và vật che khuất trong môi trường tự nhiên. Điều này khẳng định rằng, mặc dù chúng ta đã giải quyết tốt bài toán so khớp mẫu trong điều kiện kiểm soát, nhưng bài toán nhận dạng ở ngoài thực tế vẫn là

thách thức lớn cần tiếp tục giải quyết.

Trong tương lai, sự phát triển của nhận dạng dáng đi sẽ không còn đơn thuần là cuộc đua về các chỉ số độ chính xác trên tập dữ liệu cũ, mà sẽ là sự chuyển mình sang các hệ thống thông minh, bền vững và an toàn hơn. Xu hướng tất yếu sẽ là sự kết hợp đa phương thức giữa dữ liệu để vượt qua giới hạn của camera quang học, cùng với việc áp dụng các kỹ thuật học không giám sát để tận dụng nguồn dữ liệu khổng lồ chưa gán nhãn. Đồng thời, khi công nghệ này đi sâu vào đời sống, các vấn đề về bảo mật chống giả mạo và bảo tồn quyền riêng tư sẽ trở thành những trụ cột quan trọng ngang hàng với hiệu năng kỹ thuật, đảm bảo rằng nhận dạng dáng đi không chỉ là một công cụ giám sát hiệu quả mà còn là một công nghệ có trách nhiệm và đáng tin cậy.

## 2.2 Phương pháp trình bày

**Người phụ trách:** Huy.

- Trình bày chi tiết về phương pháp gốc đã chọn. (đặt vấn đề, đề xuất phương pháp, tiến hành thực nghiệm, phân tích kết quả, bàn luận, tổng kết)
- Phân tích của nhóm về hạn chế tiềm ẩn của phương pháp.

Đồ án của nhóm được phát triển dựa trên nghiên cứu **Học biểu diễn không gian-thời gian phân cấp cho nhận dạng dáng đi** [9], viết tắt là **HSTL**. Lý do nhóm chọn nghiên cứu này là:

- Được công bố năm 2023, là một trong những phương pháp SOTA tính tới thời điểm hiện tại (2025).
- Có mã nguồn mở từ chính tác giả.
- Tập dữ liệu dễ dàng truy cập.
- *Lưu ý: đề tài **Nhận diện dáng đi** có nhiều nghiên cứu được công bố kèm với mã nguồn mở, nhưng hầu hết các bộ dữ liệu cho chủ đề này đều rất lớn*

*(khoảng 100GB), không công khai mà phải yêu cầu quyền truy cập từ tác giả, đồng thời phải cung cấp nhiều thông tin không nằm trong khả năng của sinh viên trong môn học này. Cụ thể, hầu hết các tập dữ liệu bắt buộc nếu sinh viên muốn dùng, tác giả mặc định sinh viên sử dụng chúng trong các nhóm nghiên cứu của trường/khoa, nên phải có chữ ký của giảng viên, chủ nhiệm đề tài. Vì các thủ tục phức tạp và có thể tốn nhiều thời gian, gây ảnh hưởng tiến độ làm việc, nên nhóm quyết định không dùng các bộ dữ liệu giới hạn quyền truy cập kia, chỉ dùng duy nhất 1 tập công khai* ***CASIA-B*** *(trình bày sau).*

### 2.2.1 Bối cảnh và vấn đề

Các công nghệ liên quan đến sinh trắc học như vân tay, mống mắt, khuôn mặt đều yêu cầu dữ liệu được thu thập trong một điều kiện lý tưởng (ví dụ: ảnh khuôn mặt phải được chụp chính diện hoặc gần máy ảnh để xác định rõ định danh) và sự phối hợp giữa thiết bị với chủ thể (người cung cấp dữ liệu). Trong khi đó, dữ liệu dáng đi có thể được thu thập mà không cần sự phối hợp đó. Đồng thời, nghiên cứu nhận dạng dáng đi có ứng dụng trong nhiều lĩnh vực như: điều tra tội phạm, khoa học thể thao (ví dụ: phân tích chuyển động của các vận động viên), ... Tuy nhiên, các thách thức lớn dễ quan sát được, chẳng hạn sự đa dạng góc quan sát đối tượng, đối tượng bị che khuất, hoặc đối tượng mặc các trang phục sẽ làm ẩn thông tin về khuôn mẫu riêng của dáng đi.

Nhiều nghiên cứu đã được đề xuất để giải quyết các vấn đề trên. Các nghiên cứu tập trung trích xuất đặc trưng từ các loại dữ liệu như chuẩn ảnh bóng, cấu trúc cơ thể 3D, hoặc mẫu dáng đi. Dữ liệu ảnh bóng được sử dụng phổ biến nhất do dễ dàng thu thập từ các đoạn phim, đồng thời chúng cũng bảo toàn thông tin thời gian cần thiết. Một kỹ thuật phổ biến xử lý dữ liệu này là căn chỉnh và cắt ngang hình ảnh bóng. Chiến lược này lần đầu được giới thiệu trong bài toán tái nhận diện người (ReID) và đã được chứng minh là hiệu quả với nhận diện dáng đi.

Điểm hạn chế chính của kỹ thuật cắt ngang hình ảnh bóng là chúng không xem xét bản chất phân cấp tự nhiên của các chuyển động cục bộ của cơ thể người, ví dụ, chân và thân dưới có những đặc điểm chuyển động riêng biệt. Do đó, điều quan trọng là phải xem xét các vùng cơ thể này một cách riêng biệt và nghiên cứu mối quan hệ giữa các bộ phận. Đây chính là động lực nghiên cứu của **HSTL**.

**HSTL** là một quy trình học biểu diễn không gian-thời gian phân cấp, được xây dựng dựa trên 3 mô-đun chính, bao gồm:

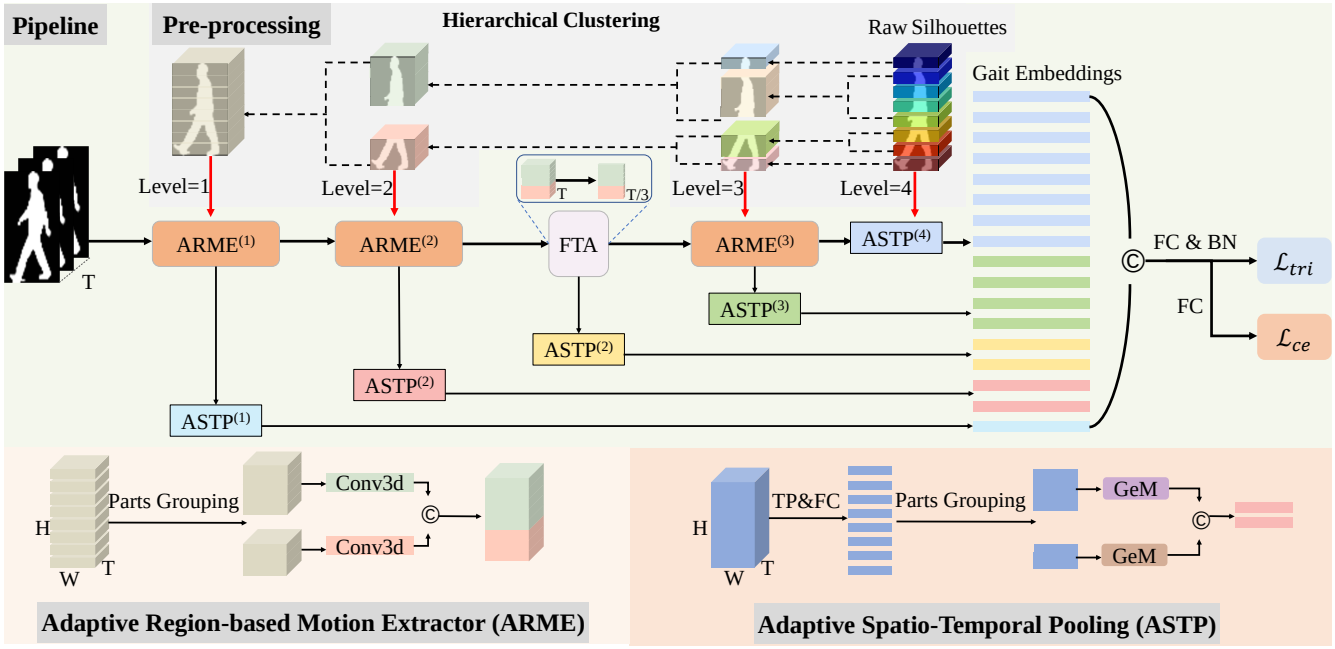
- Mô-đun trích xuất chuyển động theo vùng có khả năng thích ứng, gọi tắt là **ARME**.
- Mô-đun gộp không gian-thời gian có khả năng thích ứng, gọi tắt là **ASTP**.
- Mô-đun tổng hợp đặc trưng theo thời gian ở mức khung hình, gọi tắt là **FTA**.

**Tổng hợp các đóng góp chính của bài báo HSTL bao gồm:**

- Đề xuất quy trình đơn giản và có khả năng mở rộng, để học biểu diễn không gian-thời gian phân cấp **HSTL** cho nhận dạng dáng đi, bằng cách xem xét sự phụ thuộc các vùng trên cơ thể trong chuyển động của dáng đi.
- Giới thiệu mô-đun **ARME** để học biểu diễn không gian-thời gian độc lập với vùng cho chuỗi ảnh dáng đi, mô-đun **ASTP** để thực hiện phép ánh xạ các đặc trưng theo phân cấp, và mô-đun **FTA** để nén chuỗi dáng đi bằng cách loại bỏ cách khung hình dư thừa.
- Đạt độ hiệu quả tốt nhất (*SOTA tại thời điểm công bố* ) trên tập dữ liệu dáng đi phổ biến CASIA-B, kèm theo sự cân bằng hợp lý giữa độ phức tạp và độ chính xác của mô hình.

### 2.2.2 Phương pháp

1. Quy trình chung




Hình 4: Quy trình **HSTL**

Hình 4 trình bày quy trình **HSTL**. Cho tập dữ liệu $\mathcal{D} = \{S_i\}_{i=1}^{N}$, gồm $N$ chuỗi dáng đi, mỗi chuỗi $S_i \in \mathbb{R}^{C\times T\times H\times W}$ được biểu diễn bằng tensor 4 chiều bao gồm $C$ kênh, $T$ khung hình, và kích thước $H \times W$. Ở bước tiền xử lý, mỗi chuỗi được chia đều theo chiều ngang thành $k$ phần rồi dùng thuật toán phân cụm phân cấp (cụ thể là DBSCAN) để tạo một thứ bậc chuyển động $\mathcal{P} = \{\mathcal{P}^{(l)}\}_{l=1}^{L}$, với $L$ là số bậc trong phân cấp và $\mathcal{P}^{(l)} = \{\mathcal{P}_1^{(l)},\dots,\mathcal{P}_{K_l}^{(l)}\}$ là tập các phân vùng ở mức $l$. Sự phân cấp là một thuộc tính có cấu trúc của các mẫu chuyển động dáng đi và dùng để hướng dẫn quy trình trích xuất đặc trưng dáng đi. **HSTL** xếp chồng các mô-đun theo cấu trúc phân cấp đó. Xét đầu vào $S_{in}$, nhánh chính của **HSTL** được biểu diễn như sau:

$$Y^M = \Gamma^{(L)} \circ \Psi^{(L-1)} \circ \cdots \circ \Omega^{(2)} \circ \Psi^{(2)} \circ \Psi^{(1)}(S_{in}), \tag{7}$$

trong đó $\Psi^{(l)}$, $\Gamma^{(l)}$, và $\Omega^{(l)}$ lần lượt là các mô-đun **ARME**, **ASTP** và **FTA** ở mức thứ $l$ của $\mathcal{P}$. Vì **FTA** nén dữ liệu giữa các khung hình, nên chỉ được dùng một lần duy nhất tại mức $l_\Omega$ trong $\mathcal{P}$ (ví dụ, $l_\Omega = 2$ trong biểu thức (7))

để tránh mất thông tin quá mức.

Các đặc trưng dáng đi ở các phân cấp khác được thu thập bằng cách truyền:

- Đầu ra $Y^{(l)}$ của từng $\Psi^{(l)}$ ở mức $l \in \{1, 2, \ldots, L-1\}$,
- và đầu ra $Y_{\Omega}^{(l_\Omega)}$ của $\Omega^{(l_\Omega)}$ ở mức $l_\Omega$

vào $\Gamma^{(l)}$ tương ứng. Đầu ra cuối cùng của quy trình **HSTL** được hình thành bằng cách ghép nối các đặc trưng, biểu diễn như sau:

$$Y = \Bigg[ Y^M, \Gamma^{(L-1)}\left(Y^{(L-1)}\right), \ldots, \Gamma^{(l_\Omega)}\left(Y_{\Omega}^{(l_\Omega)}\right), \Gamma^{(2)}\left(Y^{(2)}\right), \Gamma^{(1)}\left(Y^{(1)}\right) \Bigg], \tag{8}$$

Sau đó, $Y$ được truyền qua các lớp kết nối đầy đủ (**FC**), trước khi tối ưu bằng tổ hợp hàm mất mát bộ ba $\mathcal{L}_{tri}$ và hàm mất mát entropy chéo $\mathcal{L}_{ce}$.

2. Mô-đun trích xuất chuyển động theo vùng có khả năng thích ứng - **ARME**.

   Mục tiêu của **ARME** là trích xuất các mẫu không gian-thời gian tương ứng với từng vùng cơ thể, và chúng phải độc lập với nhau. Dựa trên phân cấp $\mathcal{P}$, đầu vào $X$ được phân thành $K_l$ vùng tại mức $l$, ký hiệu là $\{X_j^{(l)}\}_{j=1}^{K_l}$, với $X_j^{(l)} \in \mathbb{R}^{C \times T \times H_j^{(l)} \times W}$ và $H_j^{(l)} = \frac{|P_j^{(l)}|}{k} H$ là chiều cao vùng $j$ ở mức $l$. Mức $l$ của **ARME**, $\Psi^{(l)}$ được biểu diễn như sauL

$$Y_{\Psi}^{(l)} = \Psi^{(l)}(X^{(l)}) = \left[ f_1(X_1^{(l)}), f_2(X_2^{(l)}), \ldots, f_{K_l}(X_{K_l}^{(l)}) \right], \tag{9}$$

   trong đó, $f_j(\cdot)$ là phép tích chập 3D (không chia sẻ) ứng với vùng $j$. Đầu ra $Y_{\Psi}^{(l)} \in \mathbb{R}^{C^{(l)} \times T \times H \times W}$ của mức $l$ có $C^{(l)}$ kênh, và giữ nguyên độ phân giải không gian và thời gian.

3. Mô-đun gộp không gian-thời gian có khả năng thích ứng - **ASTP**.

**ASTP** thực hiện ánh xạ đặc trưng phân cấp bằng cách áp dụng kĩ thuật gộp (**pooling**) không đều lên các vùng thu được từ $\mathcal{P}$. Với vùng $j$ của mức $l$, $X_j^{(l)}$, mô-đun **ASTP** tại mức $l$, $\Gamma^{(l)}$, được biểu diễn như sau:

$$Y_{\Gamma,j}^{(l)} = \Gamma_j^{(l)}(X_j^{(l)}) = \mathrm{GeM}_j \circ \mathrm{FC} \circ \mathrm{Max}(X_j^{(l)}), \tag{10}$$

trong đó, $\mathrm{Max}(\cdot)$ là phép gộp dựa trên giá trị lớn nhất (max pooling) dọc theo trục thời gian, $\mathrm{FC}(\cdot)$ là lớp kết nối đầy đủ, $\mathrm{GeM}_j(\cdot)$ là phép gộp trung bình tổng quát (generalized mean pooling (GeM)) [8] cho vùng $j$. Chuỗi phép ánh xạ của $\Gamma_j$ như sau: $\Gamma_j : \mathbb{R}^{C\times T\times H_j^{(l)}\times W} \mapsto \mathbb{R}^{C\times 1\times H_j^{(l)}\times W} \mapsto \mathbb{R}^{C^{(l)}\times 1\times H_j^{(l)}\times W} \mapsto \mathbb{R}^{C^{(l)}\times 1\times 1\times 1}$. Nối các $Y_{\Gamma,j}^{(l)}$ theo thứ tự, thu được $Y_\Gamma^{(l)} = \left[Y_{\Gamma,1}^{(l)}, Y_{\Gamma,2}^{(l)}, \ldots, Y_{\Gamma,K_l}^{(l)}\right]$, với $Y_\Gamma^{(l)} \in \mathbb{R}^{C^{(l)}\times 1\times K_l\times 1}$ là đầu ra của **ASTP** ở mức $l$.

4. Mô-đun tổng hợp đặc trưng theo thời gian ở mức khung hình - **FTA**.

   **FTA** nén các khung hình cục bộ để loại bỏ khung dư thừa bằng cách kết hợp đa tỉ lệ thời gian và chọn khung tại mức khung hình. Với vùng $X_j^{(l)}$, hai phép gộp theo thời gian với nhân $3\times 1\times 1$ và $5\times 1\times 1$ (với cùng độ trượt $3\times 1\times 1$) sinh ra $U_{j,1}^{(l)}$ và $U_{j,2}^{(l)}$, biểu diễn như sau:

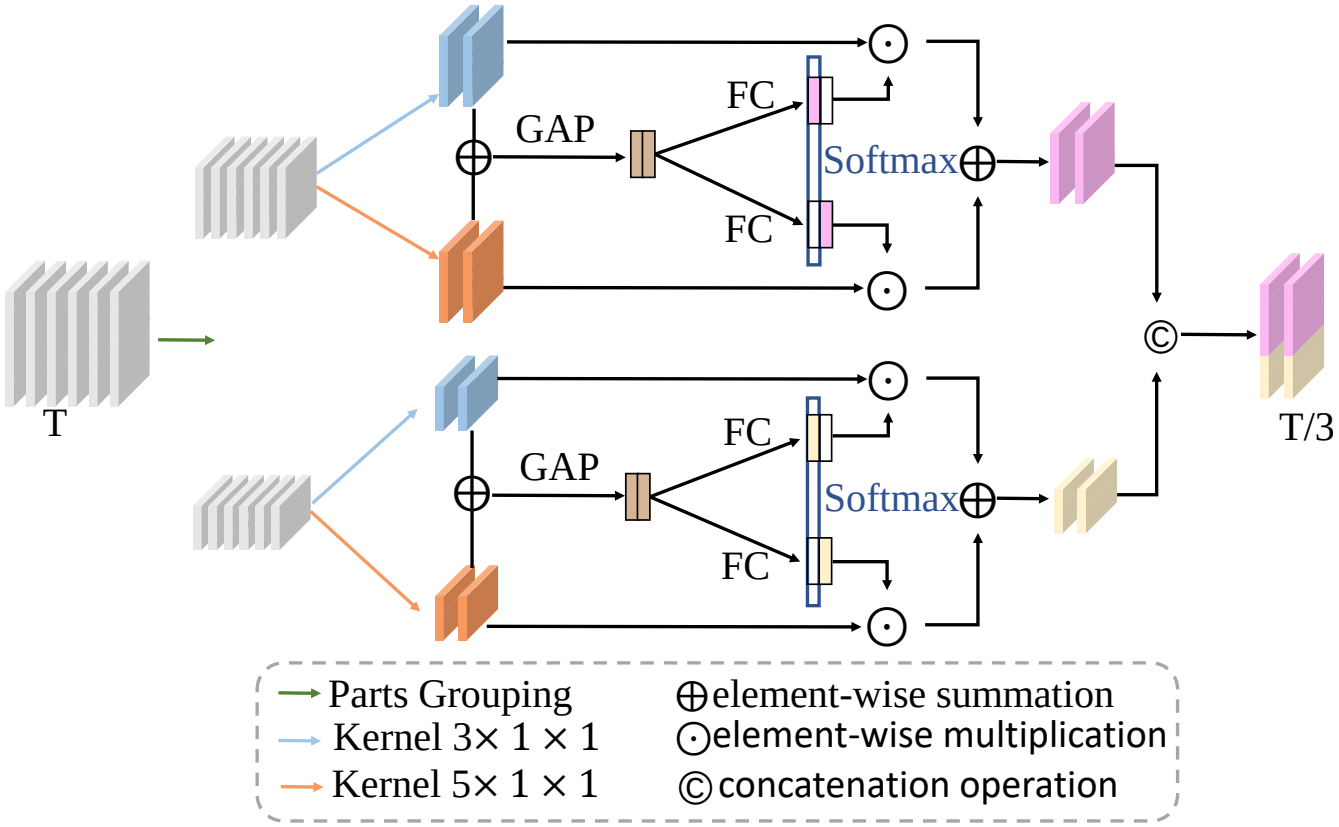
$$\begin{aligned}\hat{U}_j^{(l)} &= U_{j,1}^{(l)} + U_{j,2}^{(l)} \\ &= \mathrm{Max}_{3\times 1\times 1}^{3\times 1\times 1}\left(X_j^{(l)}\right) + \mathrm{Max}_{5\times 1\times 1}^{3\times 1\times 1}\left(X_j^{(l)}\right),\end{aligned} \tag{11}$$

   trong đó, $\mathrm{Max}_{3\times 1\times 1}^{3\times 1\times 1}(\cdot)$ và $\mathrm{Max}_{5\times 1\times 1}^{3\times 1\times 1}(\cdot)$ là phép gộp dựa trên giá trị lớn nhất. $\hat{U}_j^{(l)}$, $U_{j,1}^{(l)}$ và $U_{j,2}^{(l)}$ có cùng kích thước $(C, \frac{T}{3}, H_j^{(l)}, W)$.

   Tiếp đó, **FTA** sinh trọng số lựa trong khung:

$$\begin{aligned}Z_{j,1}^{(l)} &= \mathrm{FC}_{j,1}^{(l)}\left(\mathrm{GAP}\left(\hat{U}_j^{(l)}\right)\right), \\ Z_{j,2}^{(l)} &= \mathrm{FC}_{j,2}^{(l)}\left(\mathrm{GAP}\left(\hat{U}_j^{(l)}\right)\right),\end{aligned} \tag{12}$$

   trong đó, $\mathrm{GAP}(\cdot)$ là phép gộp trụng bình toàn cục dọc theo chiều không gian, $Z_{j,1}^{(l)}$ và $Z_{j,2}^{(l)} \in \mathbb{R}^{C\times\frac{T}{3}\times 1\times 1}$. Trọng số được chuẩn hóa chéo trên 2 tỉ lệ, được biểu




Hình 5: Minh họa mô-đun **FTA**.

diễn như sau:

$$\mathcal{W}_{j,s,c,t}^{(l)} = \frac{e^{Z_{j,s,c,t}^{(l)}}}{e^{Z_{j,1,c,t}^{(l)}} + e^{Z_{j,2,c,t}^{(l)}}} \quad s \in \{1, 2\}, \tag{13}$$

trong đó, $\mathcal{W}_{j,s,c,t}^{(l)} \in \mathbb{R}^{1\times1\times1\times1}$ là trọng số của kênh $c$của khung hình $t$. Kết hợp biểu thức (11) và biểu thức (13), đầu ra vùng $j$, $Y_{\Omega,j}^{(l)} \in \mathbb{R}^{C^{(l)}\times\frac{T}{3}\times H_j^{(l)}\times W}$ ở mức $l$ **FTA** được biểu diễn như sau:

$$Y_{\Omega,j}^{(l)} = \mathcal{W}_{j,1}^{(l)} \odot U_{j,1}^{(l)} + \mathcal{W}_{j,2}^{(l)} \odot U_{j,2}^{(l)}, \tag{14}$$

trong đó $\mathcal{W}_{j,1}^{(l)}, \mathcal{W}_{j,2}^{(l)} \in \mathbb{R}^{C\times\frac{T}{3}\times1\times1}$ là 2 tensor trọng số được tính theo biểu thức (11), và $\odot$ là phép nhân trên từng phần tử. Kết quả của **FTA**, $Y_{\Omega}^{(l)} \in \mathbb{R}^{C\times\frac{T}{3}\times H\times W}$ được tạo bằng cách ghép nối $K_l$ vùng của mức $l$, với $Y_{\Omega}^{(l)} = \left[Y_{\Omega,1}^{(l)}, Y_{\Omega,2}^{(l)}, \ldots, Y_{\Omega,K_l}^{(l)}\right]$.

### 2.2.3 Cài đặt thực nghiệm

Tác giả thực hiện trên 4 tập CASIA-B, GREW, OUVMLP, và Gait3D. Tuy nhiên, nhóm chỉ có thể tìm thấy và truy cập tập CASIA-B (lý do đã trình bày ở trên). Đây cũng là tập nhóm dùng để phát cải thiện nghiên cứu, nên nhóm tập trung trình bày thực nghiệm và kết quả trên tập này.

1. Tập dữ liệu

   CASIA-B gồm 124 chủ thể, 11 góc nhìn, ba điều kiện đi bộ: NM (normal), BG (cầm túi), CL (mặc áo khoác). Theo giao thức chuẩn, 74 chủ thể đầu dùng để huấn luyện và 50 chủ thể còn lại để kiểm thử. Trong kiểm thử, bốn chuỗi NM#01-04 được dùng làm gallery; các chuỗi còn lại (NM#05-06, BG#01-02, CL#01-02) là probe.

2. Chi tiết cài đặt thực nghiệm

   - Đầu vào: ảnh hình bóng, được các về kích thước $64 \times 44$, dùng mẫu 30 khung hình trong huấn luyện và toàn bộ khung hình trong kiểm thử.
   - Batch size $(8 \times 8)$, bộ tối ưu *Adam* với $weightdecay = 5 \times 10^{-4}$, huấn luyện 100K iterations, tốc độ học khởi tạo là $1e - 5$, và giảm 10% tại 70K.
   - Hàm mất mát bộ ba với lề $m = 0.2$.
   - Trong **GeM** của mô-đun **ASTP**, tham số $p = 6.5$.

3. Chi tiết kiến trúc của **HSTL** trong thực nghiệm với tập CASIA-B ở bảng 1.

### 2.2.4 Kết quả

Kết quả thực nghiệm trên CASIA-B với số liệu chi tiết ở bảng **??**.

Nhìn chung, **HSTL** đa phần tốt hơn các nghiên cứu pháp ở thời điểm đó. Phân tích chi tiết về bảng này sẽ được bỏ qua trong báo cáo tiến độ này để tránh dài dòng.

| Mức | Khối | Lớp | $C_{in}$ | $C_{out}$ | Kernel | $K_l$ | Các nhóm |
|---|---|---|---|---|---|---|---|
| 1 | **ARME** | Conv3d | 1 | 32 | (3,3,3) | 1 | $\{\{1,2,3,4,5,6,7,8\}\}$ |
| | **ASTP** | | | | | | |
| 2 | **ARME** | Conv3d | 32 | 32 | (3,3,3) | 2 | $\{\{1,2,3,4,5\},\{6,7,8\}\}$ |
| | | Conv3d | 32 | 64 | (3,3,3) | | |
| | **ASTP** | | | | | | |
| 2 | **FTA** | MaxPool | 64 | 64 | (3,1,1) | 2 | $\{\{1,2,3,4,5\},\{6,7,8\}\}$ |
| | | | | | (5,1,1) | | |
| | **ASTP** | | | | | | |
| 3 | **ARME** | Conv3d | 64 | 128 | (3,3,3) | 4 | $\{\{1\},\{2,3,4,5\},\{6,7\},\{8\}\}$ |
| | | Conv3d | 128 | 128 | (3,3,3) | | |
| | **ASTP** | | | | | | |
| 4 | **ASTP** | | | | | 8 | $\{\{1\},\{2\},\{3\},\{4\},$ $\{5\},\{6\},\{7\},\{8\}\}$ |

Bảng 1: Kiến trúc chi tiết của **HSTL** đề xuất trên CASIA-B. Cột đầu tiên biểu thị các mức của phân cấp dáng đi và $K_l$ là số nhóm ở mức $l$. $C_{in}$ và $C_{out}$ lần lượt biểu thị kênh đầu vào và kênh đầu ra của mỗi lớp. Các bộ phận cơ thể được đánh chỉ số theo thứ tự không gian từ trên xuống dưới, được đánh số từ 1 đến 8.

## 2.3 Hướng nghiên cứu và thực nghiệm

### 2.3.1 Thay thế kỹ thuật ARME thành P3D

**Cơ sơ sở lý thuyết**

Xét một thao tác tích chập trên một vùng cơ thể thứ $j$ tại cấp độ $l$. Giả sử đầu vào là tensor $X \in \mathbb{R}^{C_{in} \times T \times H \times W}$. Bộ lọc có kích thước không gian $k \times k$ và thời gian $d$.

Đối với kĩ thuật ARME, thì ARME sử dụng tích chập 3D tiêu chuẩn để trích xuất đặc trưng không gian và thời gian đồng thời. Cụ thể là dùng một kernel 3D kích thước $d \times k \times k$ trượt trên cả ba chiều (thời gian, chiều cao, chiều rộng).
Với mỗi vùng $j$, đầu ra $Y_j$ được tính bằng:

$$Y_j = f_j(X_j) = W_{3D} * X_j + b$$

| Gallery NM #1-4 | | 0° – 180° | | | | | | | | | | | Mean | Std |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Probe | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | | |
| NM #5-6 | GaitSet [2] | 90.8 | 97.9 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 | 3.5 |
| | GaitPart [4] | 94.1 | 98.6 | 99.3 | 98.5 | 94.0 | 92.3 | 95.9 | 98.4 | 99.2 | 97.8 | 90.4 | 96.2 | 3.1 |
| | 3D Local [6] | 96.0 | 99.0 | 99.5 | 98.9 | 97.1 | 94.2 | 96.3 | 99.0 | 98.8 | 98.5 | 95.2 | 97.5 | 1.8 |
| | CSTL [5] | 97.2 | 99.0 | 99.2 | 98.1 | 96.2 | 95.5 | 97.7 | 98.7 | 99.2 | 98.9 | 96.5 | 97.8 | 1.3 |
| | GaitGL [7] | 96.0 | 98.3 | 99.0 | 97.9 | 96.9 | 95.4 | 97.0 | 98.9 | 99.3 | 98.8 | 94.0 | 97.4 | 1.7 |
| | LagrangeGait [1] | 95.7 | 98.1 | 99.1 | 98.3 | 96.4 | 95.2 | 97.5 | 99.0 | 99.3 | 98.9 | 94.9 | 97.5 | 1.6 |
| | MetaGait [3] | 97.3 | **99.2** | 99.5 | **99.1** | 97.2 | 95.5 | 97.6 | 99.1 | 99.3 | **99.1** | 96.7 | **98.1** | 1.3 |
| | **Ours** | **97.6** | 98.0 | **99.6** | 98.2 | **97.4** | **96.5** | **97.9** | **99.3** | **99.4** | 98.4 | **97.0** | **98.1** | **1.0** |
| BG #1-2 | GaitSet [2] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 | 4.9 |
| | GaitPart [4] | 89.1 | 94.8 | 96.7 | 95.1 | 88.3 | 84.9 | 89.0 | 93.5 | 96.1 | 93.8 | 85.8 | 91.5 | 4.2 |
| | 3D Local [6] | 92.9 | 95.9 | **97.8** | 96.2 | 93.0 | 87.8 | 92.7 | 96.3 | 97.9 | 98.0 | 88.5 | 94.3 | 3.5 |
| | CSTL [5] | 91.7 | 96.5 | 97.0 | 95.4 | 90.9 | 88.0 | 91.5 | 95.8 | 97.0 | 95.5 | 90.3 | 93.6 | 3.0 |
| | GaitGL [7] | 92.6 | 96.6 | 96.8 | 95.5 | 93.5 | 89.3 | 92.2 | 96.5 | 98.2 | 96.9 | 91.5 | 94.5 | 2.8 |
| | LagrangeGait [1] | 94.2 | 96.2 | 96.8 | 95.8 | 94.3 | 89.5 | 91.7 | 96.8 | 98.0 | 97.0 | 90.9 | 94.6 | 2.7 |
| | MetaGait [3] | 92.9 | **96.7** | 97.1 | 96.4 | 94.7 | 90.4 | 92.9 | **97.2** | 98.5 | **98.1** | 92.3 | 95.2 | 2.6 |
| | **Ours** | **95.0** | 96.5 | 97.3 | **96.6** | **95.3** | **93.3** | **94.6** | 96.8 | **98.6** | 97.7 | **92.9** | **95.9** | **1.7** |
| CL #1-2 | GaitSet [2] | 61.4 | 75.4 | 80.7 | 77.3 | 72.1 | 70.1 | 71.5 | 73.5 | 73.5 | 68.4 | 50.0 | 70.4 | 8.0 |
| | GaitPart [4] | 70.7 | 85.5 | 86.9 | 83.3 | 77.1 | 72.5 | 76.9 | 82.2 | 83.8 | 80.2 | 66.5 | 78.7 | 6.6 |
| | 3D Local [6] | 78.2 | 90.2 | 92.0 | 87.1 | 83.0 | 76.8 | 83.1 | 86.6 | 86.8 | 84.1 | 70.9 | 83.7 | 6.2 |
| | CSTL [5] | 78.1 | 89.4 | 91.6 | 86.6 | 82.1 | 79.9 | 81.8 | 86.3 | 88.7 | 86.6 | 75.3 | 84.2 | 4.9 |
| | GaitGL [7] | 76.6 | 90.0 | 90.3 | 87.1 | 84.5 | 79.0 | 84.1 | 87.0 | 87.3 | 84.4 | 69.5 | 83.6 | 6.3 |
| | LagrangeGait [1] | 77.4 | 90.6 | 93.2 | 90.2 | 84.7 | 80.3 | 85.2 | 87.7 | 89.3 | 86.6 | 71.0 | 85.1 | 6.3 |
| | MetaGait [3] | 80.0 | 91.8 | 93.0 | 87.8 | 86.5 | 82.9 | 85.2 | 90.0 | 90.8 | 89.3 | 78.4 | 86.9 | **4.6** |
| | **Ours** | **82.4** | **94.2** | **95.0** | **91.7** | **88.2** | **83.3** | **88.0** | **92.3** | **93.1** | **91.0** | **78.5** | **88.9** | 5.1 |

Bảng 2: Độ chính xác Rank-1 (%) trên CASIA-B dưới tất cả các góc nhìn và các điều kiện khác nhau, loại trừ các trường hợp góc nhìn giống nhau. Std biểu thị độ lệch chuẩn mẫu hiệu suất trên 11 góc nhìn.

Trong đó $W_{3D} \in \mathbb{R}^{C_{out} \times C_{in} \times d \times k \times k}$.

Đối với kĩ thuật P3D, P3D tách một kernel 3D kích thước $d \times k \times k$ thành hai kernel riêng biệt: một kernel không gian $(1 \times k \times k)$ và một kernel thời gian $(d \times 1 \times 1)$. Theo cơ chế như sau:

- Kernel không gian $(S)$: Sử dụng bộ lọc kích thước $1 \times k \times k$, tương đương với 2D CNN để học các đặc trưng hình ảnh.
- Kernel thời gian $(T)$: Sử dụng các bộ lọc $d \times 1 \times 1$ để xây dựng các kết nối thời gian giữa các bản đồ đặc trưng liền kề.

Thay vì tính trực tiếp, ta thực hiện tuần tự:

$$Y_{Spatial} = S(X_j) = W_S * X_j \quad (\text{kernel } 1 \times k \times k)$$

$$Y_j = T(Y_{Spatial}) = W_T * Y_{Spatial} \quad (\text{kernel } d \times 1 \times 1)$$

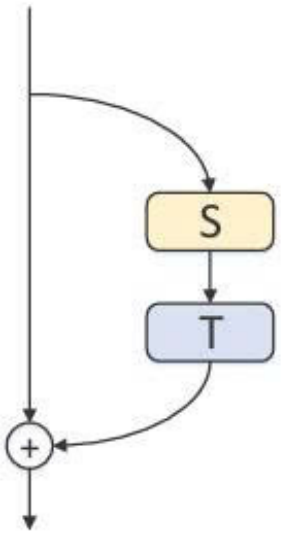
Trong đó $W_S \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times k \times k}$ và $W_T \in \mathbb{R}^{C_{out} \times C_{out} \times d \times 1 \times 1}$.

Trong nghiên cứu này, tôi áp dụng cấu trúc Residual của P3D-A. Cụ thể là thành phần thời gian (T) đi trực tiếp sau thành phần không gian (S) trên cùng một đường dẫn. Đầu ra của không gian là đầu vào của thời gian.

$$x_{t+1} = (I + T \cdot S) \cdot x_t = x_t + T(S(x_t))$$

**Ưu nhược điểm và độ phức tạp tính toán**

Về mặt tính toán, module ARME trong HSTL sử dụng tích chập 3D tiêu chuẩn $(3 \times 3 \times 3)$ nên tốn kém tài nguyên với số lượng tham số tỷ lệ thuận với $3 \times 3 \times 3 = 27$. Trong khi đó, kỹ thuật P3D tách kernel này thành hai phần riêng biệt: không gian $(1 \times 3 \times 3)$ và thời gian $(3 \times 1 \times 1)$, giúp giảm số lượng tham số xuống chỉ còn tỷ lệ với $1 \times 3 \times 3 + 3 \times 1 \times 1 = 12$. Như vậy, việc chuyển sang P3D giúp giảm khối lượng tính toán và tham số khoảng 2.25 lần. Sự tối ưu này cực kỳ quan trọng đối với kiến trúc chia vùng của HSTL, cho phép bạn xây dựng mô hình sâu hơn hoặc xử lý dữ liệu lớn hơn mà không bị quá tải bộ nhớ.




Hình 6: Kiến trúc của khối P3D-A: Các bộ lọc không gian (S) và thời gian (T) được sắp xếp nối tiếp.

Mặc dù P3D đã giảm đáng kể số lượng tham số và chi phí tính toán, tuy nhiên việc tách biệt không gian và thời gian có thể làm giảm khả năng học các mối tương quan chặt chẽ tức thời giữa hai miền này so với ARME.

**Mức độ cải thiện kỳ vọng**

Việc thay thế ARME bằng P3D được kỳ vọng sẽ mang lại các lợi ích sau:

- Giảm đáng kể chi phí tính toán và bộ nhớ, cho phép mô hình xử lý các chuỗi video dài hơn hoặc nhiều vùng cơ thể hơn mà không bị quá tải.
- Vì P3D sẽ có khả năng khái quát hóa cực tốt trên nhiều tác vụ video và bộ dữ liệu khác nhau như Sports-1M, UCF101. Khi đưa vào HSTL, nó có thể giúp mô hình hoạt động ổn định hơn trên các tập dữ liệu ngoài thực tế.
- Trong nhận diện dáng người, việc sử dụng tích chập P3D trong tập CASIA-B hy vọng sẽ cải thiện độ chính xác như trong DeepGaitV2-P3D đã cải thiện độ chính xác tới 9.1% so với phiên bản 2D thuần túy trên một số bộ dữ liệu.

### 2.3.2 Thay thế Triplet Loss thành Circle Loss

#### 1. Cơ Sở Lý Thuyết

**A. Batch-Hard Triplet Loss (HSTL Hiện Tại)** HSTL sử dụng đặc trưng dáng đi $x \in \mathbb{R}^d$ được chuẩn hóa Batch (không chuẩn hóa L2). Mục tiêu: giảm khoảng cách Euclidean $d(a, p)$ giữa mỏ neo $a$ và dương $p$, tăng $d(a, n)$ với âm $n$.

Công thức:

$$\mathcal{L}_{tri} = \sum_{(a,p,n)\in\mathcal{T}} [d(a, p) - d(a, n) + m]_+$$

Với $d(x, y) = ||x - y||_2$, $m = 0.2$.

Hạn chế: 1. Tối ưu hóa lười biếng: Gradient bằng 0 khi biên thỏa mãn, không siết chặt cụm. 2. Không gian không ràng buộc: Có thể phóng đại độ lớn đặc trưng thay vì học góc phân biệt.

**B. Circle Loss Đề Xuất** Chuẩn hóa L2 nghiêm ngặt: $x \leftarrow \frac{x}{||x||_2}$. Sử dụng độ tương tự $s_p = \cos(a, p)$, $s_n = \cos(a, n)$.

Công thức:

$$\mathcal{L}_{circle} = \log\left(1 + \sum_{n\in\mathcal{N}}\sum_{p\in\mathcal{P}} \exp\left(\gamma(\alpha_n s_n - \alpha_p s_p)\right)\right)$$

Với $\alpha_p = [O_p - s_p]_+$, $\alpha_n = [s_n - O_n]_+$, $O_p = 1 + m$, $O_n = -m$.

Lý do "Circle": Biên quyết định hình tròn trong không gian $(s_n, s_p)$, cho phép tối ưu hóa tham lam.

#### 2. Ưu/Nhược Điểm Và Độ Phức Tạp

**Ưu Điểm** 1. Tối ưu hóa tham lam: Không vùng chết, đẩy $s_p \to 1$, $s_n \to 0$ liên tục. 2. Thống nhất học cấp lớp và cặp đôi: Tăng tổng quát hóa trên góc nhìn mới. 3. Gradient thích ứng: Tăng trọng số cho mẫu khó.

**Nhược Điểm** 1. Bùng nổ gradient: Cần cân bằng ($\text{loss}_w eight : 0.05$).2.$Nhycmsiuthams$ : $Tngtc\gamma$ và $m$. 3. Mất thông tin độ lớn vector: Có thể ảnh hưởng độ tin cậy.

**Độ Phức Tạp** - Huấn luyện: Hơi cao hơn do exp/log, nhưng không đáng kể. - Suy luận: Giống hệt, dùng cosine similarity.

---

**3. Cải Thiện Dự Kiến** 1. Hội tụ nhanh: 76.3% Rank-1 (NM) tại iter 1000, nhờ tổ chức nhúng sớm. 2. Chính xác cao hơn trên CL/BG: +1-3%, do phạt phương sai nội lớp liên tục. 3. Cấu trúc hình học tốt: Nhúng trên siêu cầu, hiệu quả cho truy vấn lớn.

# 3 LỜI KẾT

Báo cáo này trình bày các nội dung tổng hợp và phân tích của nhóm về phương pháp HSTL và các hướng cải tiến đề xuất. Do hạn chế về kinh nghiệm nghiên cứu và năng lực chuyên môn, báo cáo có thể còn tồn tại các thiếu sót về mặt khoa học, cách diễn đạt chưa được tối ưu, và một số phân tích chưa thực sự làm rõ vấn đề cốt lõi. Nhóm chúng em chân thành mong nhận được các nhận xét và góp ý quý báu từ quý Thầy để có thể hoàn thiện và nâng cao chất lượng báo cáo trong giai đoạn tiếp theo.

Nhóm chúng em xin chân thành cảm ơn quý Thầy đã dành thời gian đọc và đánh giá báo cáo. Kính chúc quý Thầy và gia đình sức khỏe, hạnh phúc.

# 4 TÀI LIỆU THAM KHẢO

## Tài liệu

[1] Tianrui Chai, Annan Li, Shaoxiong Zhang, Zilong Li, and Yunhong Wang. Lagrange motion analysis and view embeddings for improved gait recognition. In *CVPR*, pages 20249–20258, 2022.

[2] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, pages 8126–8133, 2019.

[3] Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, and Xi Li. Metagait: Learning to learn an omni sample adaptive representation for gait recognition. In *ECCV*, pages 357–374. Springer, 2022.

[4] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020.

[5] Xiaohu Huang, Duowang Zhu, Hao Wang, Xinggang Wang, Bo Yang, Botao He, Wenyu Liu, and Bin Feng. Context-sensitive temporal feature learning for gait recognition. In *ICCV*, pages 12909–12918, 2021.

[6] Zhen Huang, Dixiu Xue, Xu Shen, Xinmei Tian, Houqiang Li, Jianqiang Huang, and Xian-Sheng Hua. 3d local convolutional neural networks for gait recognition. In *ICCV*, pages 14920–14929, 2021.

[7] Beibei Lin, Shunli Zhang, and Xin Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *ICCV*, pages 14648–14656, 2021.

[8] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE TPAMI*, 41(7):1655–1668, 2018.

[9] Lei Wang, Bo Liu, Fangfang Liang, and Bincheng Wang. Hierarchical spatio-temporal representation learning for gait recognition, 2023.

# 5 PHỤ LỤC

| **Tên** | **Công việc** |
|---|---|
| Phạm Thái Huy | Phương pháp trình bày (HSTL)<br>Thay thế Triplet Loss bằng Circle Loss |
| Nguyễn Đức Mạnh | Tổng hợp nội dung chương sách (phần 2, 3) |
| Lê Quang Vĩnh Quyền | Tổng hợp nội dung chương sách (phần 1, 4, 5)<br>Thay thế Conv3D bằng P3D |

Bảng 3: Bảng phân chia công việc