



VIỆN NGHIÊN CỨU DỮ LIỆU LỚN
CHƯƠNG TRÌNH ĐÀO TẠO KỸ SƯ AI

Multi-Camera Object Tracking System

Môn học: Machine Learning

Thành viên:

Lê Đăng Khoa

Phan Đình Anh Quân

Lê Tấn Lộc

Từ Cảnh Minh

Mai Việt Dũng

Ngày 12 tháng 9 năm 2024

Mục lục

1 Giới thiệu	1
1.1 Bối cảnh	1
1.2 Mục tiêu và Phạm vi Nghiên cứu	1
1.3 Thách thức	1
2 Phát biểu bài toán	2
2.1 Mô tả bài toán	2
2.2 Hướng thực hiện	2
2.2.1 Phát hiện đối tượng (Object Detection)	3
2.2.2 Theo dõi đối tượng (Object Tracking)	3
2.2.3 Nhận diện đối tượng (People Re-Identification)	4
3 Phương pháp	5
3.1 Phát hiện đối tượng (Object Detection)	5
3.2 Theo dõi đối tượng (Object Tracking)	7
3.3 Tái nhận dạng đối tượng (People Re-identification - ReID)	8
4 Thực nghiệm	10
4.1 Bộ Dữ Liệu Thực Nghiệm	10
4.1.1 Thông tin bộ dữ liệu	10
4.1.2 Tiền xử lý bộ dữ liệu	11
4.2 Chỉ số Dánh giá Object Detection	12
4.2.1 Intersection over Union (IoU)	12
4.2.2 Độ chính xác trung bình (Mean Average Precision - mAP)	13
4.3 Chỉ số Dánh giá Object Tracking	14
4.3.1 Độ chính xác phát hiện (DetA)	14
4.3.2 Độ chính xác liên kết (AssA)	15
4.3.3 MOTA (Multi-Object Tracking Accuracy)	15
4.3.4 IDF1 (Identification F1 Score)	15
4.3.5 HOTA (Higher Order Tracking Accuracy)	16
4.4 Kết quả	17

4.4.1	Kết quả Object Detection	17
4.4.2	Kết quả Object Tracking	18
4.4.3	Kết quả của bước Person ReID	19
5	Kết luận	19
	Tài liệu	20

1 Giới thiệu

1.1 Bối cảnh

Theo dõi đối tượng đa người, đa camera là một lĩnh vực nghiên cứu quan trọng trong thị giác máy tính với các ứng dụng rộng rãi và thiết yếu như giám sát an ninh, quản lý thành phố thông minh, và cải thiện khả năng phản ứng trong các tình huống khẩn cấp. Sự phát triển nhanh chóng của các hệ thống camera giám sát đã trở thành một phần không thể thiếu của hạ tầng an ninh hiện đại, cung cấp tầm quan sát toàn diện và nâng cao khả năng nhận thức tình huống trong các môi trường phức tạp. Các hệ thống này giúp cải thiện khả năng phát hiện sớm và phản ứng nhanh chóng với các sự cố, từ đó nâng cao hiệu quả quản lý và bảo vệ an ninh. Trong các thành phố lớn và khu vực công cộng, việc triển khai các hệ thống đa camera không chỉ giúp theo dõi giao thông và an ninh mà còn hỗ trợ trong các nghiên cứu xã hội và phân tích hành vi. Sự cần thiết của các giải pháp theo dõi đối tượng chính xác và hiệu quả ngày càng trở nên quan trọng khi các hệ thống giám sát ngày càng được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau.

1.2 Mục tiêu và Phạm vi Nghiên cứu

Dự án này tập trung vào việc triển khai và tối ưu hóa một mô hình theo dõi đối tượng đa người qua nhiều camera, với mục tiêu chính là hiểu rõ cách thức hoạt động và ứng dụng của mô hình trong môi trường giám sát thực tế. Thay vì áp dụng nhiều mô hình khác nhau, dự án sẽ giới hạn vào một mô hình cụ thể để tập trung nghiên cứu sâu hơn, giúp tối ưu hóa và kiểm tra hiệu suất trong việc phát hiện đối tượng và duy trì liên kết qua các khung hình liên tiếp. Mục tiêu của dự án không phải là mở rộng hay phát triển thêm về mặt lý thuyết, mà là nắm bắt các phương pháp hiện có và hiểu cách chúng có thể được triển khai hiệu quả trong thực tế. Việc giới hạn phạm vi nghiên cứu này tạo điều kiện để tích lũy kinh nghiệm và kiến thức về các yếu tố quan trọng như độ chính xác trong phát hiện đối tượng, khả năng duy trì danh tính đối tượng qua các khung hình và sự liên tục của quá trình theo dõi trong các tình huống thực tế.

1.3 Thách thức

Tuy nhiên, việc theo dõi đối tượng qua nhiều camera đem lại nhiều thách thức phức tạp hơn so với các hệ thống camera đơn lẻ. Các vấn đề chính là việc xử lý các tình huống che khuất đối

tương, sự thay đổi về ánh sáng và điều kiện môi trường giữa các camera cũng gây khó khăn trong việc duy trì tính chính xác của các thông tin.Thêm vào đó, sự khác biệt về góc nhìn và độ phân giải giữa các camera tạo ra thách thức lớn trong việc đồng bộ hóa dữ liệu và đảm bảo tính nhất quán trong việc theo dõi đối tượng. Vấn đề tái nhận dạng đối tượng khi chúng di chuyển giữa các camera khác nhau cũng là một thách thức quan trọng, đòi hỏi các phương pháp tiên tiến để phân loại và kết nối thông tin về đối tượng từ nhiều nguồn khác nhau. Các thách thức liên quan đến đồng bộ hóa camera, sự trôi dạt đặc trưng và hiệu quả tính toán cũng làm cho nhiệm vụ này trở nên phức tạp hơn, yêu cầu các giải pháp sáng tạo và hiệu quả để xử lý và đồng bộ hóa dữ liệu từ các nguồn khác nhau một cách chính xác và liên tục.

2 Phát biểu bài toán

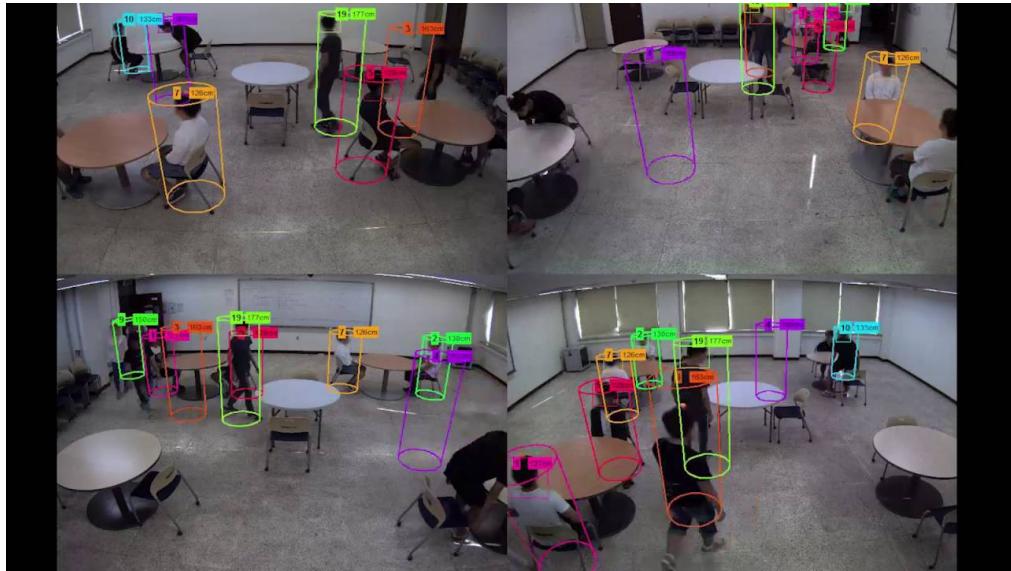
2.1 Mô tả bài toán

Dầu vào của bài toán bao gồm một loạt các khung hình video được thu thập từ các camera phân tán trong một khu vực rộng lớn. Mỗi khung hình chứa thông tin hình ảnh về các đối tượng, chủ yếu là con người, trong các cảnh vật và điều kiện ánh sáng khác nhau. Điều này tạo ra một lượng lớn dữ liệu hình ảnh với sự thay đổi về góc nhìn, ánh sáng, và bối cảnh giữa các camera.

Dầu ra của bài toán là một tập hợp thông tin chi tiết về các đối tượng xuất hiện trong các khung hình video này. Cụ thể, dầu ra bao gồm các thông tin về vị trí và danh tính của từng đối tượng qua toàn bộ hệ thống camera. Điều này bao gồm việc xác định vị trí chính xác của các đối tượng trong từng khung hình, duy trì sự nhất quán của các đối tượng qua các khung hình liên tiếp, và khả năng nhận diện lại các đối tượng khi chúng di chuyển từ camera này sang camera khác. Dầu ra cuối cùng là một tập hợp các đối tượng được theo dõi với các đường đi được vẽ ra qua các camera, cùng với các ID duy nhất hoặc nhãn để phân biệt từng đối tượng. Điều này không chỉ giúp duy trì sự liên tục của thông tin về các đối tượng mà còn hỗ trợ trong việc phân tích hành vi và các ứng dụng an ninh khác.

2.2 Hướng thực hiện

Trong quá trình xử lý và phân tích dữ liệu video từ nhiều camera, hướng thực hiện chính bao gồm ba bước quan trọng: phát hiện đối tượng, theo dõi đối tượng và nhận diện lại đối tượng.



Hình 1: Hình ảnh minh họa hệ thống Multi-Camera Multi-People Object Tracking (BMVC 2015)[2].

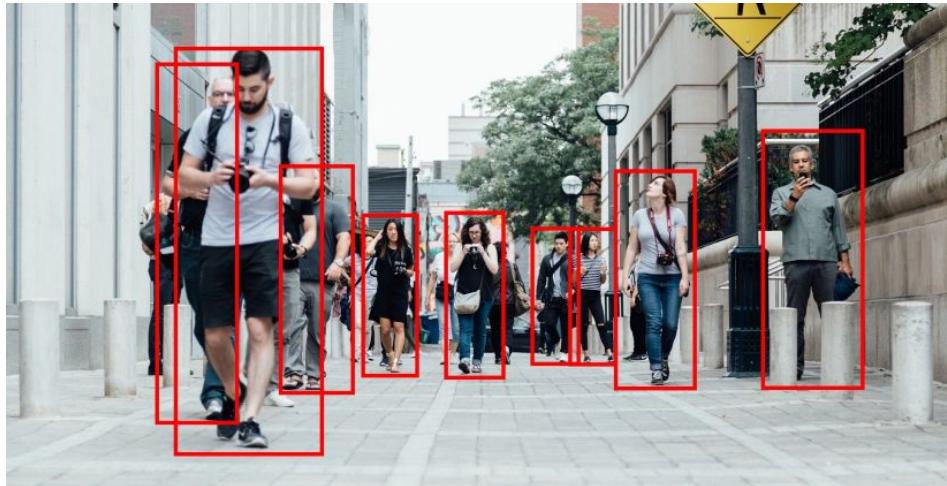
Trong quá trình xử lý và phân tích dữ liệu video từ nhiều camera, hướng thực hiện chính bao gồm hai bước quan trọng: phát hiện người (People Detection) và theo dõi người (People Tracking).

2.2.1 Phát hiện đối tượng (Object Detection)

Bước đầu tiên trong quy trình là phát hiện người, nơi chúng tôi tập trung vào việc xác định và phân loại các cá nhân trong từng khung hình video. Mục tiêu của giai đoạn này là phát hiện chính xác vị trí của từng người và xác định các thuộc tính cơ bản của họ. Quá trình này bao gồm việc phân tích các khung hình để xác định các vùng chứa người, đồng thời gán nhãn phân loại cho từng cá nhân. Kết quả của bước này là các tọa độ và nhãn cho các đối tượng người trong khung hình, giúp tạo cơ sở dữ liệu cho các bước xử lý tiếp theo.

2.2.2 Theo dõi đối tượng (Object Tracking)

Sau khi phát hiện người, bước tiếp theo là theo dõi người (People Tracking), trong đó chúng tôi sử dụng thông tin từ giai đoạn phát hiện để duy trì sự theo dõi liên tục của từng người qua nhiều khung hình video. Mục tiêu của giai đoạn này là theo dõi đường đi của từng cá nhân qua thời gian, đảm bảo rằng mỗi người được nhận diện và theo dõi liên tục mặc dù có thể di chuyển giữa các khung hình khác nhau. Bước này giúp duy trì sự nhất quán trong việc theo dõi các cá nhân qua các khung hình liên tiếp và cung cấp thông tin về hành trình di chuyển của họ trong hệ thống.



Hình 2: Phát hiện và xác định vị trí của từng người trong khung hình.

giám sát.



Hình 3: Hệ thống theo dõi đa người qua nhiều camera với các đối tượng được đánh dấu và theo dõi qua nhiều khung hình.

2.2.3 Nhận diện đối tượng (People Re-Identification)

Cuối cùng, bước nhận diện lại đối tượng (People Re-Identification) là quá trình xác định và phân loại các đối tượng khi chúng di chuyển giữa các camera khác nhau. Đây là giai đoạn quan trọng để đảm bảo rằng các đối tượng được nhận diện một cách chính xác ngay cả khi chúng xuất hiện ở các góc nhìn khác nhau hoặc trong các điều kiện ánh sáng khác nhau. Bước này giúp kết nối

thông tin về đối tượng từ các camera khác nhau, tạo ra một hệ thống theo dõi liên tục và thống nhất qua toàn bộ hệ thống giám sát.



Hình 4: Minh họa quá trình nhận diện lại đối tượng (People Re-Identification) qua các camera khác nhau.

Tổng quan, ba bước này phối hợp với nhau để cung cấp một hệ thống giám sát đa camera hiệu quả, giúp theo dõi và nhận diện đối tượng một cách chính xác qua nhiều khung hình và camera.

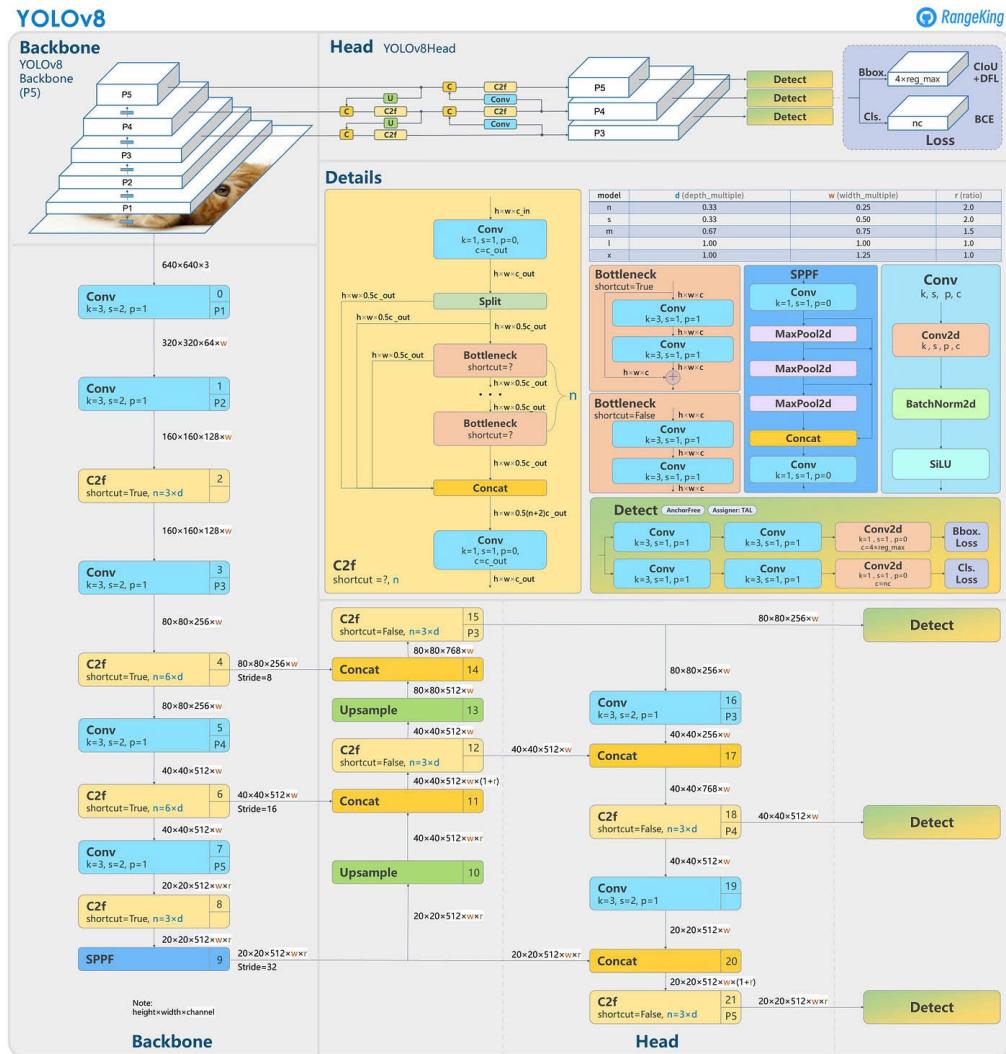
3 Phương pháp

Nhóm tiếp cận bằng phương pháp sử dụng YOLO cho việc nhận diện tất cả các đối tượng các đối tượng có trong một frame. Sau đó, dùng BoT-SORT để theo dõi các đối tượng đó qua các frame trong mỗi camera, và dùng mô hình LightMBN để tái định dạng các tracklet qua các camera và để xác định tracklet của đối tượng cần theo dõi. Trong các phần tiếp theo, nhóm sẽ trình bày chi tiết về từng module được sử dụng.

3.1 Phát hiện đối tượng (Object Detection)

Trong nghiên cứu này, chúng tôi sử dụng mô hình YOLOv8 [8] (You Only Look Once phiên bản 8) để thực hiện nhiệm vụ phát hiện đối tượng. YOLOv8 [8] là phiên bản cải tiến của họ mô hình YOLO, với tốc độ và độ chính xác vượt trội. Mô hình có khả năng phát hiện đối tượng trong thời gian thực và hoạt động hiệu quả ngay cả trong các môi trường phức tạp, như ánh sáng thay đổi, che khuất, và góc nhìn khác nhau. **Cơ chế hoạt động:**

YOLO nhận hình ảnh đầu vào với kích thước 640x640 pixels. Hình ảnh đầu vào được chia



Hình 5: Kiến trúc mô hình Yolov8[8].

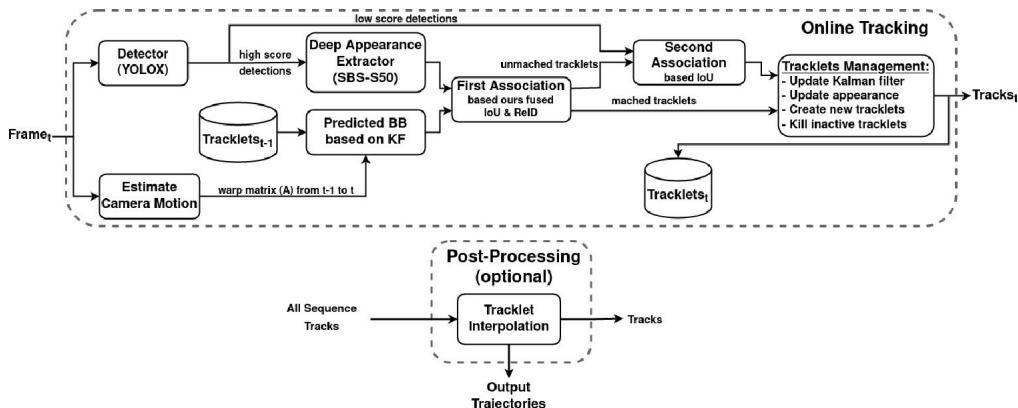
thành một lưới các ô vuông có kích thước là 19x19. Mỗi ô lưới sẽ chịu trách nhiệm phát hiện đối tượng nếu tâm của đối tượng rơi vào trong ô đó. Mỗi ô sẽ dự đoán một số lượng xác định các hộp giới hạn (bounding boxes) với 5 thông tin: tọa độ trung tâm (x, y), chiều rộng (width), chiều cao (height), và độ tin cậy (confidence score). Mỗi hộp giới hạn cũng dự đoán xác suất mà đối tượng thuộc vào mỗi lớp. Điều này có nghĩa là mỗi hộp giới hạn dự đoán một véc-tơ xác suất cho tất cả các lớp đối tượng có thể.

Sau khi tất cả các hộp giới hạn và xác suất lớp đã được dự đoán, YOLO sử dụng một bước để loại bỏ các hộp giới hạn chồng chéo quá nhiều với nhau bằng cách sử dụng Non-Maximum Suppression (NMS) để giữ lại hộp giới hạn có độ tin cậy cao nhất. Điều này giúp giảm thiểu các dự đoán dư thừa và cải thiện độ chính xác của mô hình.

3.2 Theo dõi đối tượng (Object Tracking)

Theo dõi đối tượng là công đoạn nhằm duy trì và cập nhật thông tin về vị trí và danh tính của các đối tượng qua nhiều khung hình liên tiếp. Sau khi đối tượng được phát hiện, hệ thống sẽ gán một ID duy nhất cho mỗi đối tượng và theo dõi quỹ đạo của họ theo thời gian. Việc duy trì đúng ID của đối tượng là điều đặc biệt quan trọng, nhất là trong các trường hợp đối tượng bị che khuất tạm thời hoặc có nhiều đối tượng xuất hiện đồng thời.

Để thực hiện nhiệm vụ này, chúng tôi sử dụng mô hình BoT-SORT (Boosted SORT) [1]. Đây là một phiên bản nâng cấp của SORT, được thiết kế để tăng cường độ chính xác trong việc theo dõi đối tượng. BoT-SORT không chỉ dựa trên thông tin quỹ đạo của đối tượng mà còn kết hợp các đặc trưng học sâu để đảm bảo tính nhất quán trong việc gán ID, ngay cả khi đối tượng bị che khuất hoặc có sự thay đổi về ngoại hình trong các khung hình liên tiếp.



Hình 6: Kiến trúc mô hình BoT-SORT[1].

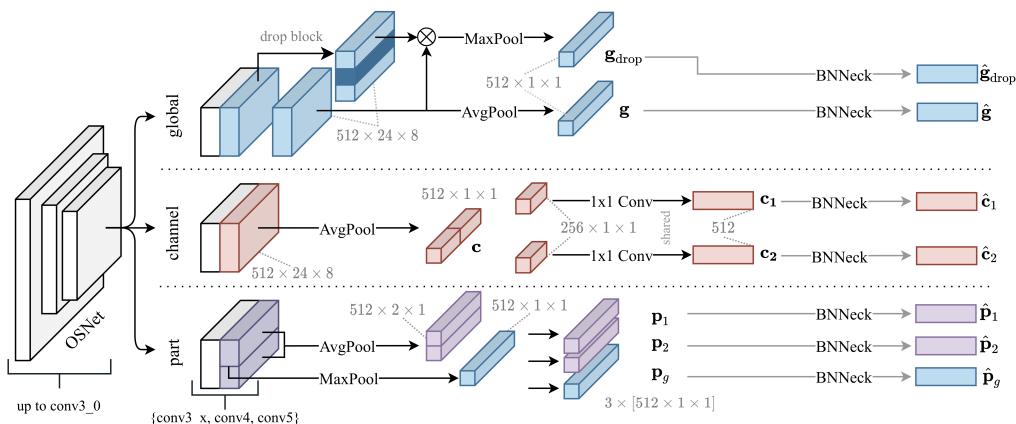
BoT-SORT bắt đầu với một mô-đun phát hiện đối tượng để nhận diện tất cả các đối tượng trong mỗi khung hình video. Sau đó, mô hình chuyển động dựa trên bộ lọc Kalman được sử dụng để dự đoán vị trí tương lai của các đối tượng. Điều này giúp hệ thống ước lượng vị trí của đối tượng trong trường hợp đối tượng bị che khuất tạm thời hoặc mất tín hiệu. Để tăng độ chính xác khi đối tượng bị che khuất hoặc có nhiều đối tượng tương tự nhau trong cùng một khung hình, BoT-SORT sử dụng một mạng nơ-ron để trích xuất các đặc trưng ngoại hình (appearance features) từ đối tượng.

Quá trình liên kết dữ liệu của BoT-SORT kết hợp cả thông tin chuyển động và đặc trưng ngoại hình. Thuật toán phân bổ Hungarian xác định đối tượng nào trong khung hình hiện tại tương ứng với đối tượng trong các khung hình trước đó. Các đối tượng bị mất tín hiệu được xử lý bằng

cách sử dụng phát hiện ngưỡng thấp từ ByteTrack, giúp khôi phục lại các đối tượng bị mất trong quá trình theo dõi.

3.3 Tái nhận dạng đối tượng (People Re-identification - ReID)

Tái nhận dạng đối tượng (ReID) là công đoạn nhận dạng lại đối tượng khi họ di chuyển giữa các camera khác nhau. Trong môi trường đa camera, đối tượng có thể xuất hiện khác nhau do sự thay đổi về góc nhìn, ánh sáng, hoặc tư thế. Do đó, việc tái nhận dạng là rất quan trọng để đảm bảo rằng hệ thống có thể duy trì danh tính chính xác của đối tượng khi họ xuất hiện ở các vùng quan sát khác nhau.



Hình 7: Kiến trúc mô hình LightMBN cho việc tái định danh đối tượng[5].

Để thực hiện nhiệm vụ tái nhận dạng, nhóm sử dụng mô hình LightMBN. Mô hình LightMBN (Light-weight Multi-Branch Network) được thiết kế với một kiến trúc mạng nơ-ron đa nhánh, nhằm tối ưu hóa cho bài toán nhận diện người (Person Re-Identification). Kiến trúc này bao gồm ba nhánh chính: nhánh toàn cục, nhánh phần, và nhánh theo kênh, mỗi nhánh đảm nhiệm một vai trò quan trọng trong việc trích xuất các đặc trưng khác nhau từ hình ảnh đầu vào.

Đầu tiên, mô hình sử dụng OSNet làm backbone cho việc trích xuất đặc trưng. Hình ảnh đầu vào được đưa qua các lớp của OSNet cho đến lớp đầu tiên của khối thứ ba, cụ thể là conv3_0. Việc lựa chọn OSNet thay vì các backbone khác như ResNet50 là do OSNet có hiệu suất vượt trội và độ phức tạp thấp hơn, điều này rất quan trọng trong các tác vụ nhận diện người, nơi mà tốc độ và độ chính xác đều cần được đảm bảo.

Sau khi xử lý qua các lớp ban đầu của OSNet, hình ảnh đầu vào X có kích thước $384 \times 128 \times 3$ sẽ được tách thành ba nhánh riêng biệt, mỗi nhánh sẽ nhận được một tensor với kích thước $24 \times 8 \times 512$.

Trong nhánh toàn cục, mô hình tạo ra hai đại diện toàn cục. Đầu tiên, thông qua việc áp dụng 2D average pooling, mô hình thu được một vector g có kích thước 512 chiều, đại diện cho thông tin tổng quát của hình ảnh. Tiếp theo, nhánh này sử dụng một drop block để loại bỏ các vùng có hoạt động cao nhất trong tensor, buộc mạng phải tập trung vào các vùng ít phân biệt hơn. Sau khi loại bỏ, 2D max pooling được áp dụng để thu được vector g_{drop} , cũng có kích thước 512 chiều, làm tăng tính bền vững của đại diện.

Tiếp theo, trong nhánh theo kênh, tensor $24 \times 8 \times 512$ sẽ được giảm xuống thành vector 512 chiều, sau đó chia thành hai vector 256 chiều. Mô hình sử dụng các convolution 1×1 để tăng quy mô trở lại, thu được hai vector 512 chiều (c_1 và c_2). Các tham số của các convolution này được chia sẻ giữa hai phần của nhánh, giúp giảm số lượng tham số tổng thể.

Trong nhánh phần, tensor ban đầu được chuyển đổi thành ba đại diện khác nhau. Đầu tiên, mô hình sử dụng average pooling để tạo ra hai đại diện phần, p_1 và p_2 , đại diện cho phần thân trên và thân dưới, tương ứng với kích thước 512 chiều. Bên cạnh đó, nhánh này cũng thu được một đại diện toàn cục p_g thông qua max pooling trên tensor ban đầu, cung cấp một cái nhìn tổng quát về hình ảnh.

Cuối cùng, mô hình sử dụng BNNeck cho tất cả các đại diện vector của từng nhánh. BNNeck bao gồm batch normalization và một lớp fully connected, có chức năng tối ưu hóa các embeddings cho hai không gian metric khác nhau: không gian xếp hạng và không gian danh tính. Các embeddings thu được trước lớp batch normalization được sử dụng để tối ưu hóa với hàm mất mát xếp hạng, trong khi các embeddings sau lớp fully connected được sử dụng cho hàm mất mát danh tính. Mô hình sử dụng hai hàm mất mát là Multi-Similarity Loss cho việc tối ưu hóa embedding vector không gian xếp hạng và Cross-Entropy Loss cho không gian phân loại danh tính. Các embeddings thu được sau lớp batch normalization nhưng trước lớp fully connected tìm kiếm một sự cân bằng giữa các đại diện của hai không gian metric khác nhau, và do đó được sử dụng cho việc suy diễn.

Kiến trúc nhẹ của LightMBN cho phép nó hoạt động hiệu quả trên các thiết bị hạn chế về tài nguyên, chẳng hạn như các thiết bị di động hoặc hệ thống nhúng. Cho thấy, LightMBN không chỉ tối ưu hóa số lượng tham số mà còn nâng cao hiệu suất trong các tác vụ nhận diện người, đặc biệt là trong các tình huống thực tế với điều kiện khác nhau.

4 Thực nghiệm

4.1 Bộ Dữ Liệu Thực Nghiệm

4.1.1 Thông tin bộ dữ liệu

Bộ dữ liệu *Large Scale Multi-Camera Detection Dataset* [9] (nguồn gốc: *The WILDTRACK Seven-Camera HD Dataset* [3]) từ Kaggle là một nguồn tài nguyên quan trọng cho nghiên cứu theo dõi đối tượng qua nhiều camera. Bộ dữ liệu này bao gồm các video ghi lại từ nhiều góc quay khác nhau trong một quảng trường lớn, với mỗi camera cung cấp một góc nhìn riêng biệt, tạo ra một cái nhìn toàn diện về hoạt động và chuyển động của các đối tượng trong khu vực này.



Hình 8: Một khung hình cắt ra từ video của camera số 1 với độ phân giải 1920 x 1080

- **Phần cứng và quá trình thu thập:** Dữ liệu này được thu thập bằng cách sử dụng 7 camera tĩnh công nghệ cao với các góc nhìn chồng chéo. Cụ thể, ba camera GoPro Hero 4 và bốn camera GoPro Hero 3 đã được sử dụng. Việc thu thập dữ liệu diễn ra trước tòa nhà chính của ETH Zurich, Thụy Sĩ, trong điều kiện thời tiết đẹp. Các chuỗi hình ảnh có độ phân giải 1920×1080 pixel, được quay ở tốc độ 60 khung hình mỗi giây.
- **Kích thước:** Bộ dữ liệu bao gồm 2800 khung hình trích xuất từ 7 camera, với 400 khung hình/camera. Mỗi khung hình có độ phân giải 1920×1080 pixel.

- **Nhận chú thích:** Mỗi frame id từ 7 camera được chú thích trong một file json tương ứng. File json này gồm số lượng từ điển tương ứng với số người khác nhau xuất hiện trong frame id này (ở 7 camera). Mỗi từ điển chứa thông tin của một người bao gồm:
 - Danh tính (personID)
 - Vị trí (positionID)
 - 7 góc nhìn, mỗi góc nhìn gồm các tọa độ trái-trên và phải-dưới của bounding box xung quanh người này
- **Thông tin hiệu chỉnh và videos:** 7 videos, mỗi video trong khoảng 30 - 40 phút

4.1.2 Tiền xử lý bộ dữ liệu

Để đảm bảo bộ dữ liệu đồng nhất với yêu cầu đầu vào của mô hình YOLOv8, các bước tiền xử lý sau được thực hiện:

- Trích xuất thông tin từ các file json dưới dạng: <object class> <object id> X Y W D
 - <object class> = 0 là pedestrian
 - <object id> phân biệt giữa những người khác nhau
 - $X, Y \in (0, 1)$ là tọa độ tâm của bounding box
 - $W, H \in (0, 1)$ là kích thước bounding box theo phần trăm của kích cỡ ảnh
- Phân chia tập dữ liệu:
 - Training: 80% (320 frames mỗi camera)
 - Validation: 10% (40 frames mỗi camera)
 - Testing: 10% (40 frames mỗi camera)
- Cấu hình file wildtrack.yaml quy định đường dẫn tới các tập train, val, test và số object class (trong project này là 1)

```

    ▼ "root" : [ 38 items
      ▼ 0 : { 3 items
        "personID" : int 122
        "positionID" : int 456826
        ▼ "views" : [ 7 items
          ▼ 0 : { 5 items
            "viewNum" : int 0
            "xmax" : int 1561
            "xmin" : int 1510
            "ymax" : int 299
            "ymin" : int 139
          }
          ▶ 1 : {....} 5 items
          ▶ 2 : {....} 5 items
          ▶ 3 : {....} 5 items
          ▶ 4 : {....} 5 items
          ▶ 5 : {....} 5 items
          ▶ 6 : {....} 5 items
        ]
      ]
    ]
  ]
}

```

Hình 9: Thông tin về một người trong khung hình đầu tiên với định dạng json: personID (định danh), positionID (vị trí), góc quay từ các camera (views) bao gồm tọa độ bounding box xung quanh người đó

4.2 Chỉ số Đánh giá Object Detection

4.2.1 Intersection over Union (IoU)

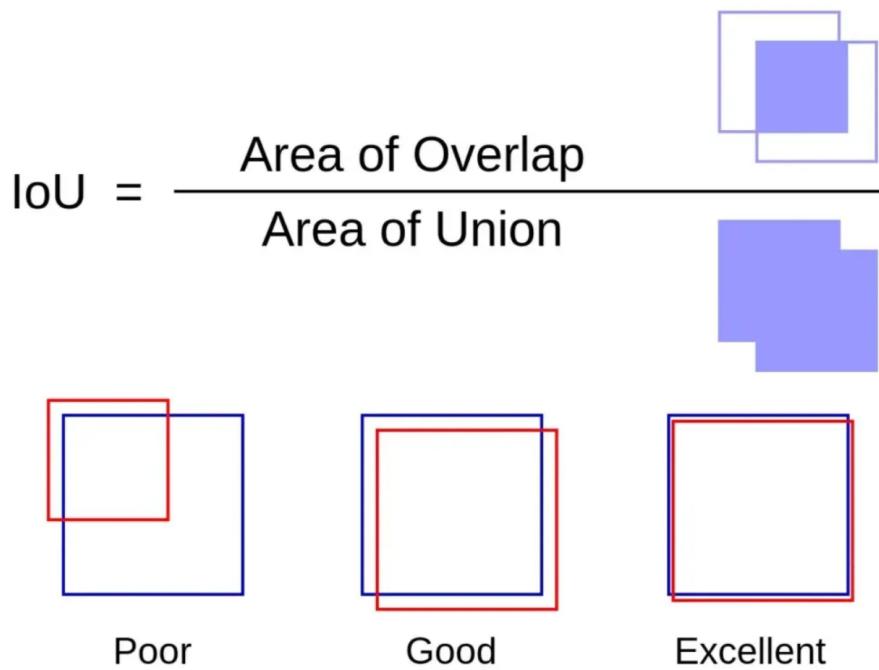
Intersection over Union (IoU) là chỉ số cơ bản để đo lường mức độ trùng lắp giữa hộp dự đoán (predicted bounding box) và hộp thực (ground truth bounding box). IoU tính bằng tỷ lệ giữa diện tích giao nhau của hai hộp và diện tích hợp nhất của chúng. Công thức tính IoU là:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

Trong đó:

- **Area of Intersection:** Diện tích giao nhau của hai hộp.
- **Area of Union:** Diện tích hợp nhất của hai hộp.

IoU có giá trị từ 0 đến 1, với giá trị gần 1 cho thấy sự trùng khớp cao giữa dự đoán và thực tế, trong khi giá trị gần 0 cho thấy ít hoặc không trùng lắp. Một ngưỡng IoU (thường là 0.5) được sử dụng để xác định tính chính xác của dự đoán.



Hình 10: Một sơ đồ minh họa chỉ số Intersection over Union (IoU) [4]

4.2.2 Độ chính xác trung bình (Mean Average Precision - mAP)

Dộ chính xác trung bình (Mean Average Precision - mAP) là chỉ số quan trọng để đánh giá hiệu suất của hệ thống phát hiện đối tượng. Để tính toán mAP, ta thực hiện các bước sau:

Trước tiên, tính Average Precision (AP) cho từng lớp đối tượng. AP được tính bằng cách lấy diện tích dưới đường Precision-Recall (PR) curve. Đường PR được xác định từ độ chính xác (Precision) và độ nhạy (Recall) tại các ngưỡng khác nhau của điểm số dự đoán.

Công thức tính Precision và Recall như sau:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Trong đó:

- TP (True Positives) là số đối tượng được phát hiện đúng.
- FP (False Positives) là số đối tượng bị phát hiện sai là có mặt.

- ***FN*** (False Negatives) là số đối tượng không được phát hiện nhưng thực tế có mặt.

Đường PR được xây dựng từ các giá trị Precision và Recall ở các ngưỡng khác nhau của điểm số dự đoán. Diện tích dưới đường PR là giá trị AP cho một lớp. Công thức tính AP là:

$$AP = \int_0^1 \text{Precision}(r) dr$$

Sau khi tính AP cho tất cả các lớp, mAP được tính bằng trung bình cộng của các giá trị AP:

$$mAP = \frac{1}{C} \sum_{c=1}^C AP_c$$

Trong đó, C là số lớp đối tượng và AP_c là AP của lớp c . mAP cung cấp cái nhìn tổng quan về khả năng phát hiện của mô hình trong việc nhận diện nhiều lớp đối tượng. Một giá trị mAP cao cho thấy hiệu suất phát hiện tốt hơn, trong khi giá trị thấp chỉ ra hiệu suất kém hơn.

4.3 Chỉ số Đánh giá Object Tracking

Chúng tôi đánh giá hệ thống theo dõi đối tượng bằng cách sử dụng một số chỉ số chuẩn, tập trung vào hai khía cạnh chính là độ chính xác phát hiện (DetA) và độ chính xác liên kết (AssA). Sau đó, các chỉ số MOTA, IDF1 và HOTA được sử dụng để cân bằng và đưa ra cái nhìn tổng quan về hiệu suất của hệ thống.

4.3.1 Độ chính xác phát hiện (DetA)

Độ chính xác phát hiện (Detection Accuracy - DetA) đo lường mức độ chính xác trong việc phát hiện các đối tượng trong một cảnh. Công thức tính DetA là:

$$\text{DetA} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

Trong đó:

- **TP** (True Positives): Số lượng đối tượng phát hiện đúng.
- **FP** (False Positives): Số lượng đối tượng bị phát hiện sai.
- **FN** (False Negatives): Số lượng đối tượng không được phát hiện.

4.3.2 Độ chính xác liên kết (AssA)

Độ chính xác liên kết (Association Accuracy - AssA) đo lường khả năng duy trì liên kết đúng giữa các đối tượng qua các khung hình. Công thức tính AssA là:

$$\text{AssA} = \frac{\text{Correct Associations}}{\text{Total Associations}}$$

Trong đó:

- **Correct Associations:** Số lượng liên kết đúng giữa các đối tượng qua các khung hình.
- **Total Associations:** Tổng số liên kết dự kiến.

4.3.3 MOTA (Multi-Object Tracking Accuracy)

MOTA[6] (Multi-Object Tracking Accuracy) là chỉ số tổng hợp đo lường hiệu suất theo dõi bằng cách kết hợp các lỗi phát hiện, theo dõi và nhận dạng. Công thức tính MOTA là:

$$\text{MOTA} = 1 - \frac{\text{FN} + \text{FP} + \text{ID Sw}}{\text{Total Objects}}$$

Trong đó:

- **FP** (False Positives): Số lượng đối tượng bị phát hiện sai.
- **ID Sw** (Identity Switches): Số lượng nhầm lẫn về danh tính đối tượng.
- **Total Objects:** Tổng số đối tượng thực tế.

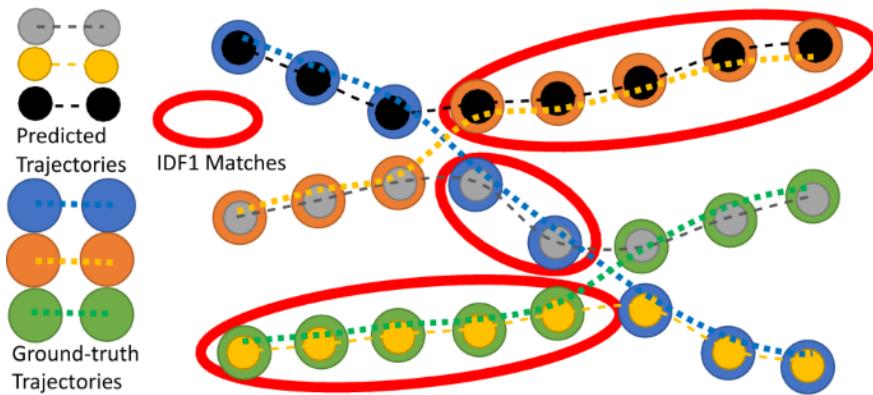
4.3.4 IDF1 (Identification F1 Score)

IDF1 (Identification F1 Score) là chỉ số kết hợp giữa độ chính xác (Precision) và độ bao phủ (Recall) trong việc nhận dạng đối tượng. Công thức tính IDF1 là:

$$\text{IDF1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong đó:

- **Precision:** Tỷ lệ đối tượng được phát hiện đúng trên tổng số đối tượng được phát hiện.
- **Recall:** Tỷ lệ đối tượng được phát hiện đúng trên tổng số đối tượng thực tế.



Hình 11: Một sơ đồ minh họa chỉ số Identification F1 Score

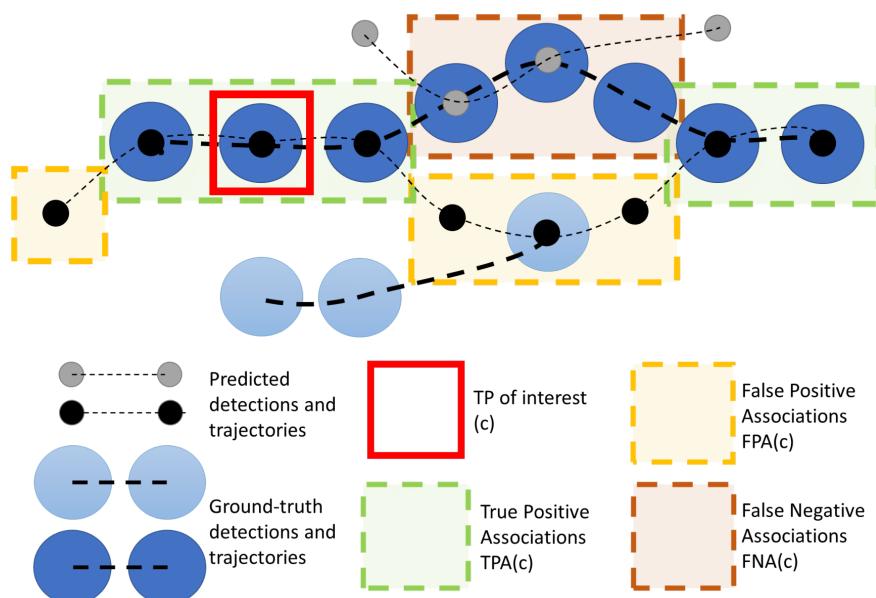
4.3.5 HOTA (Higher Order Tracking Accuracy)

HOTA^[7] (Higher Order Tracking Accuracy) là chỉ số tổng hợp mới, kết hợp độ chính xác phát hiện và độ chính xác liên kết. Công thức tính HOTA là:

$$\text{HOTA} = \sqrt{\text{DetA} \times \text{AssA}}$$

Trong đó:

- **DetA:** Độ chính xác phát hiện.
- **AssA:** Độ chính xác liên kết.



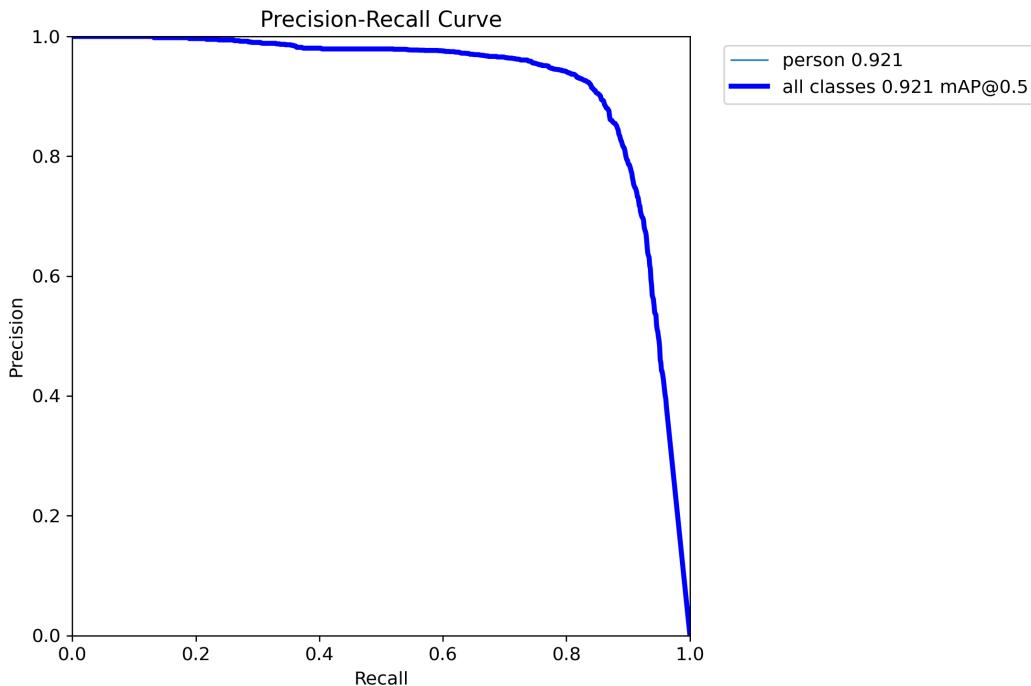
Hình 12: Một sơ đồ minh họa chỉ số Higher Order Tracking Accuracy [7]

Bằng cách sử dụng các chỉ số này, chúng tôi có thể đánh giá chi tiết và toàn diện về hiệu suất của hệ thống theo dõi đối tượng, từ khả năng phát hiện và liên kết đến khả năng duy trì nhận dạng chính xác của các đối tượng trong môi trường đa camera.

4.4 Kết quả

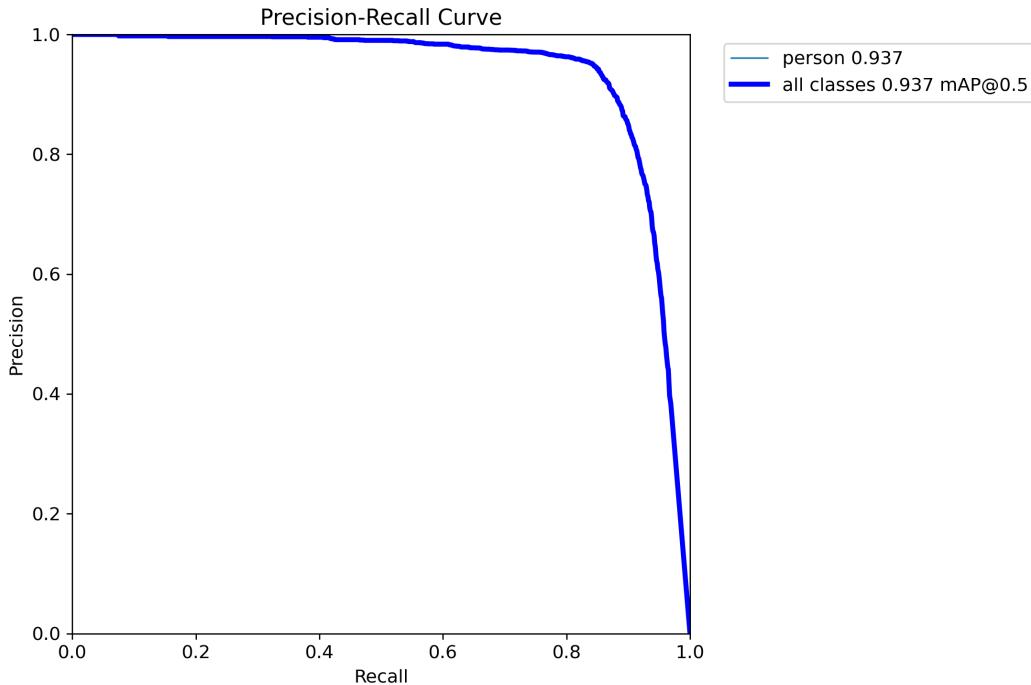
4.4.1 Kết quả Object Detection

Biểu đồ Precision-Recall được sử dụng để đánh giá hiệu suất phát hiện đối tượng trên hai tập dữ liệu: tập kiểm thử và tập xác thực. Kết quả cho thấy mô hình đạt được hiệu suất rất cao trên cả hai tập dữ liệu, với độ chính xác và độ bao phủ đều đồng đều. Trên **tập kiểm thử**, mAP@0.5 (Mean



Hình 13: Biểu đồ Precision-Recall trên tập kiểm thử với mAP@0.5 đạt 0.921.

Average Precision tại ngưỡng IoU 0.5) đạt giá trị 0.921. Kết quả này cho thấy mô hình có khả năng phát hiện đối tượng "person" với độ chính xác cao. Biểu đồ Precision-Recall thể hiện rõ sự duy trì độ chính xác (Precision) gần như hoàn hảo ở các mức Recall thấp và trung bình, và chỉ giảm nhẹ khi Recall đạt giá trị cao hơn. Điều này phản ánh rằng mô hình hoạt động rất tốt trên các tình huống phổ biến, có khả năng xử lý hầu hết các đối tượng một cách chính xác, ngay cả khi gặp các trường hợp khó khăn hơn như đối tượng bị che khuất hoặc nhỏ trong khung hình.



Hình 14: Biểu đồ Precision-Recall trên tập xác thực với mAP@0.5 đạt 0.937.

Tiếp theo, trên **tập xác thực**, mô hình tiếp tục đạt được hiệu suất ấn tượng với mAP@0.5 đạt giá trị 0.937. Precision duy trì gần 1.0 cho đến khi Recall đạt gần mức tối đa, tương tự như trên tập kiểm thử. Điều này cho thấy mô hình có sự ổn định trong việc phát hiện đối tượng trên nhiều tình huống và khung hình khác nhau. Kết quả trên tập xác thực cũng cho thấy sự ít xuất hiện của các phát hiện sai (False Positives), khi Precision chỉ giảm nhẹ ở các mức Recall rất cao.

Như vậy, trên cả hai tập kiểm thử và tập xác thực, mô hình phát hiện đối tượng đạt được hiệu suất vượt trội với mAP cao hơn 0.9, thể hiện khả năng phát hiện đối tượng với độ chính xác cao và khả năng xử lý tốt các tình huống phức tạp. Kết quả này là minh chứng cho sự hiệu quả của mô hình trong việc phát hiện đối tượng "person", và cung cấp cơ sở vững chắc cho việc áp dụng trong các ứng dụng thực tế như giám sát và theo dõi đối tượng.

4.4.2 Kết quả Object Tracking

Bảng dưới đây cung cấp kết quả của mô hình MPNTrack.

Mô hình	DetA	AssA	MOTA	IDF1	HOTA
BoT-SORT (Ours)	54.822	28.525	51.792	44.351	48.07

Bảng 1: Thông số mô hình trên tập dataset WILDTRACK

Thuật toán huấn luyện trên tập dataset WILDTRACK đạt kết quả ở mức trung bình, tuy chưa bằng các kết quả tốt nhất đã đạt được trên các mô hình SOTA.

4.4.3 Kết quả của bước Person ReID

Bảng 2: Kết quả ReID có và không có Pretrained Weights

Pretrained Weight	mAP	Top-1 Accuracy	Top-5 Accuracy	Top-10 Accuracy
No (not using pretrained)	0.3590	0.5991	0.7670	0.8394
Yes (using pretrained)	0.3937	0.6570	0.7974	0.8336

Khi không sử dụng pretrained weights, kết quả mAP đạt 0.3590, Rank1 đạt 0.5991, Rank3 đạt 0.7236, Rank5 đạt 0.7670 và Rank10 đạt 0.8394. Hiệu suất này cho thấy mô hình đạt hiệu quả thấp hơn khi không sử dụng pretrained weights, đặc biệt là ở các chỉ số quan trọng như mAP và Rank1, điều này có thể gây ảnh hưởng đến khả năng chẩn đoán.

Ngược lại, khi sử dụng pretrained weights, kết quả mAP tăng lên 0.3937, Rank1 đạt 0.6570, Rank3 đạt 0.7612, Rank5 đạt 0.7974 và Rank10 đạt 0.8336. Thời gian huấn luyện trong trường hợp này cũng ngắn hơn, chỉ mất 120 phút 13 giây. Việc sử dụng pretrained weights giúp cải thiện hiệu suất mô hình đáng kể so với khi không sử dụng, với mAP tăng 9.68% và Rank1 tăng 9.66%. Điều này cho thấy việc sử dụng pretrained weights là một giải pháp hữu ích cho việc nâng cao độ chính xác của mô hình.

5 Kết luận

Qua bài báo cáo này, nhóm nghiên cứu đã giới thiệu một cách tổng quan về bài toán theo dõi đối tượng qua nhiều camera. Nhóm đã trình bày phương pháp tiếp cận của mình, trong đó sử dụng mô hình YOLO để nhận diện đối tượng, thuật toán BoT-SORT để theo dõi đối tượng trong từng camera, và mô hình LightMBN để tái định dạng đối tượng qua nhiều camera. Các mô hình này đã thể hiện kết quả khả quan khi được đánh giá trên bộ dữ liệu WILDTRACK, tuy nhiên vẫn cần cải thiện thêm trong tương lai. Trong các bước tiếp theo, nhóm sẽ phát triển một thuật toán có khả năng nhận diện đồng thời nhiều đối tượng trên nhiều camera và nâng cao độ chính xác của các mô hình.

Tài liệu

- [1] Niv Aharon, Alex Bronstein, and Miki Korman. BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2, 2022.
- [2] Object Tracking AI. Multi-camera multi-target tracking (bmvc 2015). <https://www.youtube.com/watch?v=dliRQ9zOFPU>, 2022. Accessed: September 12, 2024.
- [3] Tsvetelina Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, César Jose, Louis Lettry, Pascal Fua, and François Fleuret. The wildtrack multi-camera person dataset. *arXiv preprint arXiv:1707.09299*, 2017.
- [4] Idiot Developer. What is intersection over union (iou)? <https://idiotdeveloper.com/what-is-intersection-over-union-iou/>, 2023. Accessed: September 12, 2024.
- [5] Fabian Herzog, Xunbo Ji, Torben Teepe, Stefan Hörmann, Johannes Gilg, and Gerhard Rigoll. Lightweight multi-branch network for person re-identification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1129–1133, 2021.
- [6] VisAI Labs. Evaluating multiple object tracking accuracy and performance metrics in a real-time setting. <https://visailabs.com/evaluating-multiple-object-tracking-accuracy-and-performance-metrics-in-a-real-time-setting>, 2023. Accessed: September 12, 2024.
- [7] Jonathon Luiten, Aljosa Osep, Laura Leal-Taixe, Bastian Leibe, Amin Ahmadi, and Tat-Jun Chin. Hota: A higher order metric for evaluating multi-object tracking. <https://arxiv.org/pdf/2009.07736.pdf>, 2020. Accessed: September 12, 2024.
- [8] Dillon Reis. Real-time flying object detection with yolov8. <https://arxiv.org/pdf/2305.09972.pdf>, 2024. Accessed: 2024-05-22.
- [9] Arya Shah. Large scale multicamera detection dataset. <https://www.kaggle.com/datasets/aryashah2k/large-scale-multicamera-detection-dataset>, 2021. Accessed: 2024-09-12.