ÁP DỤNG STRONGSORT VÀO DEEPSORT
STRONGSORT: MAKE DEEPSORT GREAT AGAIN

Trịnh Viết Khải

Khoa Công Nghệ Thông Tin, Trường Đại học Kinh Tế Tài Chính thành phố Hồ Chí Minh

Tóm tắt: *Gần đây, theo dõi đa đối tượng (MOT - Multi-object Tracking) nhận được sự quan tâm lớn và đạt được một số thành tựu đáng kể. Tuy nhiên, phương pháp tồn tại có xu hướng sử dụng đa dạng các mô hình cơ bản (vd: mô hình phát hiện và nhúng - detector and embedding models) và các thủ thuật huấn luyện khác nhau. Do đó, việc xây dựng một cơ sở tốt để so sánh là điều cần thiết. Trong bài viết này, phương pháp cổ điển DeepSORT được xem lại và cải thiện đáng kể từ nhiều khía cạnh như phát hiện đối tượng (object tracking) , nhúng (embedding) và liên kết quỹ đạo (trajectory association). Đặc biệt, không giống như hầu hết các phương pháp liên kết các đoạn*

đường ngắn thành các quỹ đạo hoàn chỉnh với tốc độ phức tạp tính toán cao, ngược lại nhóm tác giả đã đề xuất mô hình liên kết không phụ thuộc vào ngoại hình AFLink (Appearance-free Link) để thực hiện liên kết toàn cầu mà không cần thông tin về ngoại hình, và đạt được cân bằng về tốc độ và độ chính xác. Xa hơn nữa, nhóm tác giả đã đề xuất GSI (Gaussian-smoothed interpolation) dựa trên hồi quy quy trình Gaussian để giảm bớt việc phát hiện thiếu. AFLink và GSI có thể dễ dàng tích hợp vào nhiều trình theo dõi khác nhau với chi phí tăng thêm nhỏ (1.7ms và 7.1ms mỗi ảnh trên MOT17). Cuối cùng, bằng cách kết hợp StrongSORT và AFLink và GSI, công cụ theo dõi StrongSORT++ đạt được tiến triển rất tốt trên các tập dữ liệu như MOT17, MOT20, DanceTracking và KITTI`.

Keywords:
Trong bài viết:
- MOT (Multi-object Tracking - Theo dõi đa đối tượng),

- LI (Linear Interpolation - Nội suy tuyến tính),

- GSI (Gaussian-smoothed interpolation - Nội suy làm mịn Gaussian),

- AFLink (Apperance-free Link - Liên kết không phụ thuộc ngoại hình),

- GPR (Gaussian process regression - Hồi quy quy trình Gaussian),

- TBD (Tracking By Detection - Theo dõi bởi sự phát hiện)
Trong chương V:
- ID-based F1 score - sử dụng để đánh giá độ chính xác của quá trình liên kết các đối tượng theo dõi dựa trên ID của chúng.

- HOTA (Higher-order Tracking Accuracy) - là một chỉ số tổng hợp để đánh giá hiệu suất của hệ thống theo dõi đối tượng dựa trên cả việc phát hiện và liên kết các đối tượng.

- MOTA (Multiple Object Tracking Accuracy) - đánh giá độ chính xác của hệ thống theo dõi đối tượng, tập trung vào việc đo lường số lượng sự nhầm lẫn và bỏ sót các đối tượng trong quá trình theo dõi.

- FPS (Frame Per Second) - sử dụng để đo lường tốc độ xử lý và hiệu suất của các phiên bản theo dõi.

- MC (Matching Cascade) - thuật toán Matching Cascade.

- woC (abandoning Matching Cascade) - thử nghiệm để kiểm tra hiệu quả của công cụ theo dõi về việc bỏ qua Matching Cascade.

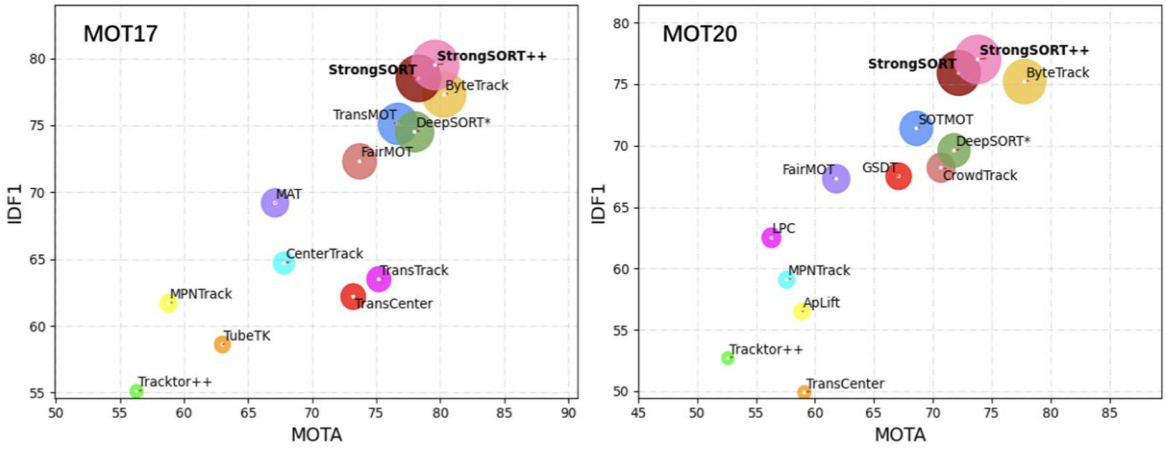
1. Giới Thiệu

Theo dõi đa đối tượng (MOT) nhắm tới phát hiện và theo dõi cụ thể các đối tượng theo từng khung hình, đóng vai trò quan trọng trong việc hiểu video. Trong những năm gần đây, mô hình TBD chi phối toàn bộ MOT và các phương pháp TBD thực hiện phát hiện trên mỗi khung hình, xây dựng MOT như một nhiệm vụ liên kết dữ liệu. Dựa trên phương pháp cổ điển TBD tracker DeepSORT, được cải tiến với mô hình nhúng và trình phát hiện mạnh mẽ kèm theo thủ thuật suy luận từ các công trình gần đây giúp tạo ra 1 phương pháp có tên là StrongSORT và đạt được kết quả tiên tiến trên các tiêu chuẩn phổ biến, bao gồm MOT17 và MOT20. Nhóm tác giả cũng đã đề cập đến hai vấn đề trong MOT (missing association and missing detection) thiếu liên kết và phát hiện thiếu:

Đối với Missing Association - liên kết thiếu, các mô hình hiện nay sử dụng liên kết toàn cầu được đề xuất để liên kết các đoạn đường ngắn thành các quỹ đạo sử dụng thông tin toàn cầu. Ngược lại, nhóm tác giả giới thiệu một mô hình liên kết không phụ thuộc ngoại hình (AFLink) chỉ sử dụng thông tin không gian và thời gian để liên kết, đạt được sự cân bằng tốt hơn giữa tốc độ và độ chính xác.

Về Missing Detection - phát hiện thiếu, nội suy tuyến tính khá phổ biến do tính đơn giản của nó, tuy nhiên độ chính xác của nó bị hạn chế vì không sử dụng thông tin chuyển động, thay vào đó nhóm tác giả đã đề xuất thuật toán nội suy làm mịn GSI - sử dụng hồi quy quy trình Gaussian để sửa các hộp giới hạn được nội suy, cải thiện độ chính xác của các vị trí được nội suy.

Cả AFLink và GSI đều là các mô hình nhẹ, độc lập và không có hình thức bên ngoài, được nhóm tác giả áp dụng để nâng cao StrongSORT và các công cụ theo dõi tiên tiến khác, với chi phí tính toán không đáng kể . Các thử nghiệm mở rộng chứng minh tính hiệu quả của các phương pháp này và khi áp dụng cho StrongSORT, chúng tạo ra một công cụ theo dõi mạnh



mẽ hơn có tên là StrongSORT++ giúp đạt được hiệu suất SOTA trên nhiều điểm chuẩn khác nhau.

2. Các Công Việc Liên Quan:

2.1. Phương Pháp Theo Dõi Chung Và Riêng Biệt

Các phương pháp theo dõi của MOT có thể phân loại thành các bộ theo dõi riêng biệt và chung. Các trình theo dõi riêng biệt tuân theo mô hình theo dõi phát hiện, mô hình này định vị mục tiêu trước rồi liên kết chúng với thông tin về hình dáng, chuyển động, v.v.. Mặt khác,

các phương pháp theo dõi chung đã được đề xuất để cùng huấn luyện phát hiện và các thành phần khác, chẳng hạn như mô hình chuyển động, nhúng và liên kết. Những bộ theo dõi chung này cung cấp hiệu suất tương đương với chi phí tính toán thấp.

Bên cạnh đó, một số nghiên cứu gần đây đề xuất loại bỏ các đặc điểm ngoại hình và chỉ dựa vào hiệu suất cao của máy dò và thông tin chuyển động, để đạt được tốc độ và hiệu suất cao hơn. Tuy nhiên, việc loại bỏ các đặc điểm ngoại hình này có thể dẫn đến sự không ổn định trong các tình huống phức tạp hơn. Trong bài viết này, nhóm tác giả đã áp dụng mô hình DeepSORT và kết hợp các kỹ thuật tiên tiến để xác nhận tính hiệu quả của khuôn khổ cổ điển này.

2.2. **Liên Kết Toàn Cầu Trong Theo Dõi Đa Đối Tượng:**

Thiếu liên kết đặt ra một thách thức đáng kể trong các nhiệm vụ của MOT. Đã có một số phương pháp tinh chỉnh kết quả theo dõi bằng mô hình liên kết toàn cầu, các phương pháp này tạo ra các bản tracklet chính xác nhưng không đầy đủ dựa trên thông tin về không gian và/hoặc hình thức, sau đó liên kết chúng bằng cách sử dụng thông tin toàn cầu theo cách ngoại tuyến. Nhiều thuật toán khác nhau đã được đề xuất, chẳng hạn như TNT, TPM, ReMOT và GIAOTracker, dựa vào các đặc điểm ngoại hình nhưng mang lại chi phí tính toán cao. Tuy nhiên, mô hình AFLink đề xuất chỉ khai thác thông tin chuyển động, dự đoán độ tin cậy liên kết giữa các tracklet. AFLink mang lại lợi ích cho các trình theo dõi hiện đại với chi phí tăng thêm không đáng kể, khiến nó trở thành một lựa chọn đơn giản và nhẹ hơn.

2.3. **Nội Suy Trong Theo Dõi Đa Đối Tượng:**

Nội suy tuyến tính thường được sử dụng để lấp đầy các khoảng trống trong quỹ đạo được phục hồi để phát hiện thiếu. Tuy nhiên, phép nội suy tuyến tính bỏ qua thông tin chuyển động, hạn chế độ chính xác của các giới hạn được khôi phục. Một số chiến lược đã được đề xuất để sử dụng hiệu quả các thông tin không gian thời gian trong nội suy như V-IOUTracker mở rộng IOUTracker bằng cách quay trở lại theo dõi một đối tượng khi xảy ra phát hiện thiếu. MAT áp dụng chiến lược lấp đầy quỹ đạo giả quan sát theo chu kỳ để làm trơn các quỹ đạo được nội suy tuyến tính. Các phương pháp khác như mô hình CMC và bộ lọc Kalman được sử dụng để dự đoán các vị trí bị thiếu.

Mặt khác, nhóm tác giả đã sử dụng mô hình GSI, mô hình chuyển động phi tuyến dựa trên thuật toán hồi quy quá trình Gaussian để làm mịn các tracklet chưa được nội suy để dự đoán vận tốc chính xác. GSI sử dụng thuật toán GPR để làm mịn đạt được sự cân bằng tốt giữa độ chính xác và hiệu quả mà không cần thêm mô hình.

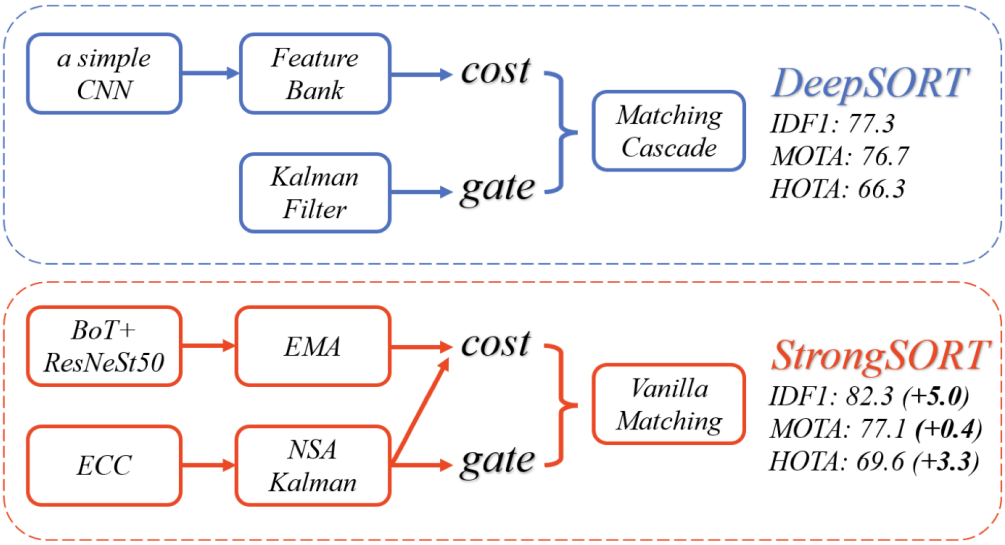
3. **StrongSORT**

3.1. **Đánh giá về DeepSORT**

Tại phần 3, nhóm tác giả giới thiệu StrongSORT, là một phiên bản nâng cấp hơn của DeepSORT. DeepSORT được mô tả như một khung hai nhánh bao gồm một nhánh ngoại hình và một nhánh chuyển động (appearance branch and motion branch) được thể hiện ở hình bên dưới.

Trong nhánh ngoại hình, bộ mô tả ngoại hình sâu (deep appearance descriptor) sẽ trích xuất các đặc điểm ngoại hình từ các phát hiện bằng mô hình CNN đã được huấn luyện trước. Cơ

chế Feature Bank được sử dụng để lưu trữ các tính năng của 100 khung hình cuối cùng cho mỗi



tracklet. Khi có các phát hiện mới, khoảng cách cosine nhỏ nhất giữa tracklet i và detection j là:

$$d(i,j) = min\{1 - f_j^T f_k^{(i)} \mid f_k^{(i)} \in B_i\}.$$

Trong nhánh chuyển động sử dụng thuật toán lọc Kalman để dự đoán vị trí của các tracklet trong khung hiện tại. Nó hoạt động theo quy trình hai giai đoạn, tức là dự đoán trạng thái và cập nhật trạng thái. Trong đó, $\hat{x}'_k$ và P'_k là trạng thái tại bước thời gian k:

$$\hat{x}'_k = F_k \hat{x}_{k-1},$$

$$P'_k = F_k P_{k-1} F_k^T + Q_k,$$

Kalman gain được tính dựa trên hiệp phương sai của trạng thái ước lượng Pk và quan sát tiếng ồn R_k, H_k^Tlà mô hình quan sát.

$$K = P'_k H_k^T (H_k P'_k H_k^T + R_k)^{-1},$$

Sau đó Kalman Gain K được sử dụng để cập nhật trạng thái cuối cùng:

$$x_k = \hat{x}'_k + K(z_k - H_k \hat{x}'_k),$$

$$P_k = (I - K H_k) P'_k,$$

Dựa vào trạng thái chuyển động của các trackets và phát hiện mới, khoảng cách Mahalanobis được sử dụng để đo lường sự khác biệt về không gian, thời gian. Cuối cùng, DeepSORT lấy khoảng cách chuyển động này để lọc ra các mối liên kết khó xảy ra.

Cuối cùng là sử dụng thuật toán Matching Cascade để giải quyết nhiệm vụ liên kết dưới dạng một loạt các bài toán con thay vì bài toán gán toàn cục.Ý tưởng là ưu tiên kết hợp tốt hơn cho các đối tượng được nhìn thấy thường xuyên hơn.

3.2. **StrongSORT**

Nhóm tác giả cũng giới thiệu những tiến bộ trong StrongSORT được cải tiến từ DeepSORT. Bao gồm việc sử dụng các mô-đun nâng cao như YOLOX-X làm máy

dò và BoT làm công cụ trích xuất tính năng ngoại hình mạnh mẽ hơn, giúp trích xuất nhiều đặc điểm phân biệt hơn. Cơ chế Feature Bank được thay thế bằng chiến lược cập nhật tính năng dựa trên đường trung bình di chuyển theo cấp số nhân để giảm độ nhạy đối với nhiễu phát hiện (**EMA**).

$$e_i^t = \alpha e_i^{t-1} + (1-\alpha) f_i^t,$$

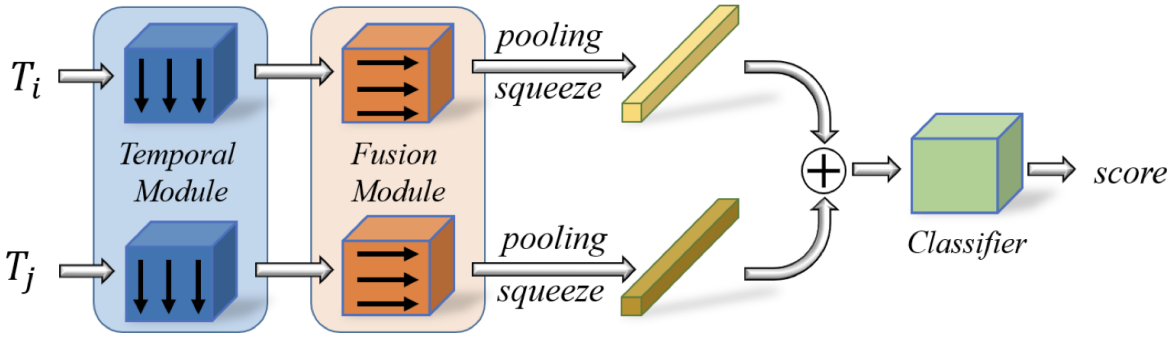
Mô hình tối đa hóa hệ số tương quan nâng cao (**ECC**) được áp dụng để bù chuyển động của máy ảnh. Đây là một kỹ thuật căn chỉnh hình ảnh tham số có thể ước tính độ xoay và dịch chuyển toàn cục giữa các khung liền kề.

$$E_{ECC}(\mathbf{p}) = \left\| \frac{\bar{\mathbf{i}}_\mathbf{r}}{\|\bar{\mathbf{i}}_\mathbf{r}\|} - \frac{\bar{\mathbf{i}}_\mathbf{w}(\mathbf{p})}{\|\bar{\mathbf{i}}_\mathbf{w}(\mathbf{p})\|} \right\|^2,$$

Bộ lọc vanilla Kalman với chất lượng thấp và bỏ qua thông tin về thang đo nhiễu phát hiện. Để giải quyết vấn đề này, nhóm tác giả đã sử dụng thuật toán **NSA Kalman** được mượn từ GIAOTracker và sử dụng để tính toán một cách thích ứng phương sai nhiễu dựa trên điểm phát hiện, cải thiện độ chính xác của các trạng thái được cập nhật. Thuật toán này đề xuất công thức tính hiệp phương sai nhiễu $\widetilde{R}_k$:

$$\widetilde{R}_k = (1-c_k)R_k,$$

Chi phí chuyển động - **Motion Cost.** DeepSORT chỉ sử dụng khoảng cách đối tượng xuất hiện làm chi phí phù hợp trong giai đoạn liên kết đầu tiên, trong đó khoảng cách chuyển động chỉ được sử dụng làm cổng. Thay vào đó, nhóm tác giả đã giải bài toán gán với cả thông tin về hình



thức và chuyển động. Ma trận chi phí C là tổng trọng số của chi phí xuất hiện Aa và chi phí chuyển động Am với λ = 0.98 như sau:

$$C = \lambda A_a + (1-\lambda) A_m,$$

Vanilla Matching - Mặc dù Matching Cascade của DeepSORT vẫn phù hợp nhưng khi trình theo dõi trở nên phức tạp hơn, điều đó sẽ khiến MC dễ gây nhầm lẫn cho các liên kết. Do đó việc ràng buộc bổ sung trước đó sẽ hạn chế độ chính xác của việc so khớp. Cuối cùng, nhóm tác giả đã quyết định sử dụng Vanilla Matching cho StrongSORT, thay thế thuật toán Matching Cascade bằng phép gán tuyến tính toàn cục vanilla.

4. StrongSORT++

Trong Phần I của bài báo, nhóm tác giả đã giới thiệu hai thuật toán nhẹ, độc lập với mô hình được gọi là AFLink và GSI để giải quyết các thách thức về việc thiếu liên kết và phát hiện thiếu trong theo dõi nhiều đối tượng (MOT). Các thuật toán này sau đó được tích hợp vào phương thức cuối cùng có tên StrongSORT++.

4.1 AFLink

AFLink được thiết kế như một mô hình không phụ thuộc ngoại hình để dự đoán khả năng kết nối giữa hai tracklet chỉ bằng thông tin không gian và thời gian.
Hình bên dưới cho thấy khung 2 nhánh của mô hình AFLink, lấy hai tracklet Tj và Ti làm đầu vào. Trong đó Mô-đun thời gian (Temporal Module) được sử dụng để trích xuất các tính năng bằng cách tích chập dọc theo chiều thời gian, sau đó Mô-đun kết hợp (Fusion Module) được sử dụng để tích hợp thông tin từ các kích thước đặc trưng khác nhau. Sau đó hai kết quả đặc tính được kết hợp và nén thành các vector đặc trưng. Cuối cùng, một perceptron nhiều lớp (MLP) được sử dụng để dự đoán điểm tin cậy cho sự liên kết.

Trong quá trình đào tạo, thủ tục liên kết được đóng khung như một nhiệm vụ phân loại nhị phân và được tối ưu hóa với tổn thất entropy chéo nhị phân như sau:

$$L_n^{BCE} = -(y_n log(\frac{e^{x_n}}{e^{x_n} + e^{1-x_n}}) + (1 - y_n) log(1 - \frac{e^{1-x_n}}{e^{x_n} + e^{1-x_n}})),$$

Trong quá trình liên kết, nhóm tác giả đã nhóm ra các cặp tracklet không hợp lý được lọc ra bằng cách sử dụng các ràng buộc về không gian và thời gian, sau đó liên kết toàn cục được giải quyết dưới dạng nhiệm vụ gán tuyến tính với khả năng kết nối điểm dự đoán.

4.2 GSI

GSI được đề xuất như một thuật toán nội suy nhẹ để lấp đầy các khoảng trống trong quỹ đạo do phát hiện thiếu. Một số chiến lược đã được đưa ra để giải quyết vấn đề này như Mô-đun Time-consuming, vd: single object-tracker, Kalman-filter hay ECC.Thay vào đó, nhóm tác giả đã sử dụng thuật toán GSI dựa trên hồi quy quy trình Gaussian để mô hình hóa chuyển động phi tuyến.

Nhóm tác giả đã xây dựng mô hình GSI cho quỹ đạo thứ i như sau:

$$p_t = f^{(i)}(t) + \epsilon,$$

với t ∈ f là frame id, pt ∈ P là biến toạ độ tại frame t, cho các quỹ đạo được theo dõi và nội suy tuyến tính (i), nhiệm vụ mô hình hoá chuyển động phi tuyến được giải quyết bằng cách khớp hàm $f^{(i)}$, nhóm tác giả đã giả sử tuân theo quy luật Gaussian:

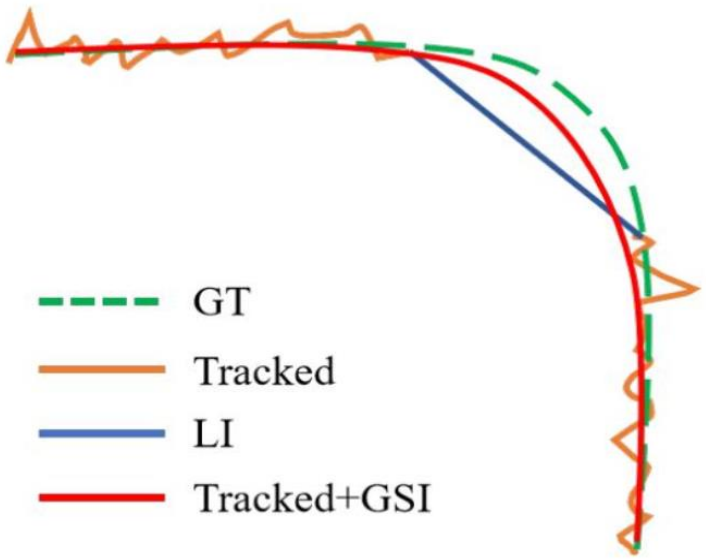
$$f^{(i)} \in GP(0, k(\cdot, \cdot)),$$

Trên cơ sở các tính chất của quá trình Gaussian, cho một bộ frame F mới, với vị trí P được làm mịn dự đoán bởi:

$$P^* = K(F^*, F)(K(F, F) + \sigma^2 I)^{-1} P,$$

Hơn nữa, siêu tham số λ kiểm soát độ trơn của quỹ đạo, nhóm tác giả đã thiết kế ra một hàm thích ứng như sau:

$$\lambda = \tau * log(\tau^3/l),$$



Minh hoạ sự khác biệt giữa LI và GSI

Hình trên minh hoạ về sự khác biệt giữa GSI và phép nội suy tuyến tính (LI). Các kết quả được theo dõi thô (màu cam) thường bao gồm nhiễu và LI (màu xanh) bỏ qua thông tin chuyển động. Nhưng với GSI (màu đỏ) của nhóm tác giả giải quyết đồng thời cả 2 vấn đề bằng cách làm trơn toàn bộ quỹ đạo với hệ số mịn và thuật toán thích ứng.

Bằng cách tích hợp AFLink và GSI với phương pháp StrongSORT, công cụ theo dõi cuối cùng, StrongSORT++, đạt được kết quả tiên tiến trên nhiều điểm chuẩn công khai, bao gồm MOT17, MOT20, DanceTrack và KITTI.

5. Thí Nghiệm:

5.1 Cài Đặt:

Bộ dữ liệu: Nhóm tác giả đã tiến hành thử nghiệm trên tập dữ liệu MOT17 là bộ dữ liệu chính, đi cùng là tập dữ liệu MOT20 bao gồm các cảnh với mật độ cao hơn. Bên cạh đó nhóm tác giả cũng thử nghiệm StrongSORT++ trên KITTI và DanceTrack, KITTI là bộ dữ liệu phổ biến bao gồm nhiều cảnh người đi bộ và ô tô, liên quan đến nhiệm vụ lái xe tự động. DanceTrack là tập dữ liệu được đề xuất gần đây để theo dõi nhiều người, khuyến khích MOT ít phụ thuộc vào hình ảnh và phụ thuộc nhiều hơn vào phân tích chuyển động.

Số liệu: Nhóm tác giả sử dụng các số liệu MOTA, ID, IDF1, HOTA, AssA, DetA và FPS để đánh giá hiệu suất theo dõi.

Chi tiết thực hiện: Chi tiết triển khai hệ thống như sau:

- ***Detection***: Thuật toán YOLOX-X được sử dụng làm máy dò, tạo ra sự cân bằng giữa thời gian và độ chính xác. Lịch trình đào tạo tuân theo cách tiếp cận tương tự như công việc trước đó. Trong quá trình suy luận, ngưỡng triệt tiêu không tối đa là 0,8 được đặt cùng với ngưỡng tin cậy phát hiện là 0,6.

- ***StrongSORT***: Các tham số cho StrongSORT được định cấu hình như sau: ngưỡng khoảng cách khớp được đặt thành 0,45, chế độ cong vênh cho ECC là MOTION EUCLIDEAN, thuật ngữ động lượng (α) trong EMA (Trung bình di chuyển theo cấp số nhân) được đặt thành 0,9 và hệ số trọng lượng (λ) cho chi phí ngoại hình được đặt thành 0,98.
- ***GSI*** (Nội suy đồng bộ hóa khoảng cách): Khoảng cách tối đa được phép cho phép nội suy là 20 khung hình và siêu tham số (τ) được đặt thành 10.
- ***AFLink***: Mô-đun tạm thời (Temoral Module) của AFLink bao gồm bốn lớp tích chập với các hạt nhân 7x1 và các kênh đầu ra là {32, 64, 128, 256}. Mỗi tích chập được theo sau bởi lớp BN (Chuẩn hóa hàng loạt) và lớp kích hoạt ReLU. Mô-đun tổng hợp bao gồm tích chập 1x3, lớp BN và lớp ReLU mà không thay đổi số lượng kênh. Bộ phân loại là MLP (Perceptron nhiều lớp) với hai lớp được kết nối đầy đủ và lớp ReLU ở giữa.
- ***Tạo dữ liệu huấn luyện***: Các quỹ đạo có chú thích được chia thành các tracklet bằng cách thêm nhiễu không gian-thời gian ngẫu nhiên, với tỷ lệ 1:3 cho các mẫu dương tính và âm tính.
- ***Tối ưu hóa và đào tạo***: Trình tối ưu hóa Adam được sử dụng cùng với chức năng mất entropy chéo. Mô hình được đào tạo trong 20 kỷ nguyên bằng cách sử dụng lịch trình tốc độ học tập ủ cosine, với quá trình đào tạo tổng thể mất hơn 10 giây.
- ***Suy luận***: Ngưỡng khoảng cách tạm thời là 30 khung hình và ngưỡng khoảng cách không gian là 75 pixel được áp dụng để lọc ra các cặp liên kết không hợp lý. Một mối liên quan được xem xét nếu điểm dự đoán của nó cao hơn 0,95.
- ***Phần cứng***: Tất cả các thử nghiệm được thực hiện trên một máy chủ có GPU V100 duy nhất.

5.2 **Nghiên Cứu Loại Bỏ**:

Nghiên cứu loại bỏ cho StrongSORT.

Bảng I tóm tắt đường dẫn từ DeepSORT đến StrongSORT:

1) BoT: Thay thế trình trích xuất tính năng ban đầu bằng BoT dẫn đến sự cải thiện đáng kể cho IDF1 (+2.2).

2) ECC: Mô hình CMC dẫn đến IDF1 (+0,2) và MOTA (+0,3 tăng nhẹ), ngụ ý rằng nó giúp trích xuất thông tin chuyển động chính xác hơn.

3) NSA: Bộ lọc NSA Kalman cải thiện HOTA (+0,4) nhưng không cải thiện MOTA và IDF1. Điều này có nghĩa là nó tăng cường độ chính xác định vị.

4) EMA: Cơ chế cập nhật tính năng EMA không chỉ mang lại khả năng liên kết vượt trội (+0,4 IDF1) mà còn mang lại tốc độ nhanh hơn (+1,2 FPS).

TABLE I: Ablation study on the MOT17 validation set for basic strategies, i.e., stronger feature extractor (BoT), camera motion compensation (ECC), NSA Kalman filter (NSA), EMA feature updating mechanism (EMA), matching with motion cost (MC) and abandoning matching cascade (woC). (best in bold)

Method	BoT	ECC	NSA	EMA	MC	woC	IDF1(↑)	MOTA(↑)	HOTA(↑)	FPS(↑)
Baseline	-	-	-	-	-	-	77.3	76.7	66.3	**13.8**
StrongSORTv1	✓						79.5	76.8	67.8	8.3
StrongSORTv2	✓	✓					79.7	77.1	67.9	6.3
StrongSORTv3	✓	✓	✓				79.7	77.1	68.3	6.2
StrongSORTv4	✓	✓	✓	✓			80.1	77.0	68.2	7.4
StrongSORTv5	✓	✓	✓	✓	✓		80.9	77.0	68.9	7.4
StrongSORTv6	✓	✓	✓	✓	✓	✓	**82.3**	**77.1**	**69.6**	7.5

TABLE II: Results of applying AFLink and GSI to various MOT methods. All experiments are performed on the MOT17 validation set with a single GPU. (best in bold)

Method	AFLink	GSI	IDF1(↑)	MOTA(↑)	HOTA(↑)	FPS(↑)
StrongSORTv1	-	-	79.5	76.8	67.8	**8.3**
	✓		80.0	76.8	68.1	8.2
	✓	✓	**80.4(+0.9)**	**78.2(+1.4)**	**68.9(+1.1)**	7.8 (-0.5)
StrongSORTv3	-	-	79.7	77.1	68.3	**6.2**
	✓		80.5	77.1	68.6	6.1
	✓	✓	**80.9(+1.2)**	**78.7(+1.6)**	**69.5(+1.2)**	5.9 (-0.3)
StrongSORTv6	-	-	82.3	77.1	69.6	**7.5**
	✓		82.5	77.1	69.6	7.4
	✓	✓	**83.3(+1.0)**	**78.7(+1.6)**	**70.8(+1.2)**	7.0 (-0.5)
CenterTrack [66]	-	-	64.6	66.8	55.3	**14.4**
	✓		68.3	66.9	57.2	14.1
	✓	✓	**68.4(+3.8)**	**66.9(+0.1)**	**57.6(+2.3)**	12.8 (-1.6)
TransTrack [45]	-	-	68.6	67.7	58.1	**5.8**
	✓		69.1	67.7	58.3	5.8
	✓	✓	**69.9(+1.3)**	**69.6(1.9)**	**59.4(+1.3)**	5.6 (-0.2)
FairMOT [64]	-	-	72.7	69.1	57.3	**12.0**
	✓		73.2	69.2	57.6	11.8
	✓	✓	**74.2(+1.5)**	**71.1(+2.0)**	**59.0(+1.7)**	10.9 (-1.1)

5) MC: Phù hợp với cả liên kết hỗ trợ ngoại hình và chi phí chuyển động (+0,8 IDF1).
6) woC: Đối với trình theo dõi mạnh hơn, thuật toán Matching Cascade với thông tin dư thừa trước đó sẽ hạn chế độ chính xác của việc theo dõi. Chỉ cần sử dụng phương pháp so khớp đơn giản, IDF1 được cải thiện với biên độ lớn (+1,4).

Nghiên cứu loại bỏ cho AFLink và GSI:

Nhìn vào bảng II, nhóm tác giả đã áp dụng AFLink và GSI cho 6 công cụ khác nhau gồm 3 phiên bản StrongSORT và 3 công cụ theo dõi hiện đại (CenterTrack, TransTrack và FairMOT) (Kết quả được hiển thị dưới bảng II). Dòng đầu tiên thể hiện các phương pháp, ứng dụng của AFLink (dòng thứ 2) mang lại mức độ cải thiện khác nhau cho các công cụ the dõi. Cụ thể, những công cụ theo dõi kém hơn có xu hướng hưởng lợi nhiều hơn từ AFLink do thiếu liên kết. Dòng thứ 3 cho kết quả mỗi công cụ theo dõi chứng minh tính hiệu quả của GSI đối với cả phát hiện và liên kết (detection and association). Khác với AFLink, GSI hoạt động tốt hơn so với các trình theo dõi mạnh hơn, nhưng nó cũng có thể nhầm lẫn bởi số lượng lớn liên kết sai với công cụ kém hơn.

Nghiên cứu loại bỏ cho Vanilla Matching:

Nhóm tác giả trình bày sự so sánh giữa 2 thuật toán Matching Cascade và Vanilla

Matching tại bảng III. Nhận thấy rằng Matching Cascade mang lại lợi ích rất lớn cho DeepSORT.

Tuy nhiên, với sự cải tiến dần dần của trình theo dõi, Matching Cascade ngày càng có những lợi thế nhỏ hơn và thậm chí có hại cho độ chính xác của việc theo dõi. Cụ thể, đối với

TABLE III: Ablation study on the MOT17 validation set for the matching cascade algorithm and vanilla matching.

Method	Matching	IDF1(↑)	MOTA(↑)
DeepSORT	Cascade	77.3	76.7
	Vanilla	76.2 (**-1.1**)	76.7 (**-0.0**)
StrongSORTv1	Cascade	79.5	76.8
	Vanilla	79.6 (**+0.1**)	76.7 (**-0.1**)
StrongSORTv2	Cascade	79.7	77.1
	Vanilla	79.7 (**+0.0**)	77.1 (**+0.0**)
StrongSORTv3	Cascade	79.7	77.1
	Vanilla	79.9 (**+0.2**)	77.1 (**+0.0**)
StrongSORTv4	Cascade	80.1	77.0
	Vanilla	81.9 (**+1.8**)	76.9 (**-0.1**)
StrongSORTv5	Cascade	80.9	77.0
	Vanilla	82.3 (**+1.4**)	77.1 (**+0.1**)

StrongSORTv5, nó có thể mang lại mức tăng 1,4 trên IDF1 bằng cách thay thế Matching Cascade với Vanilla Matching.

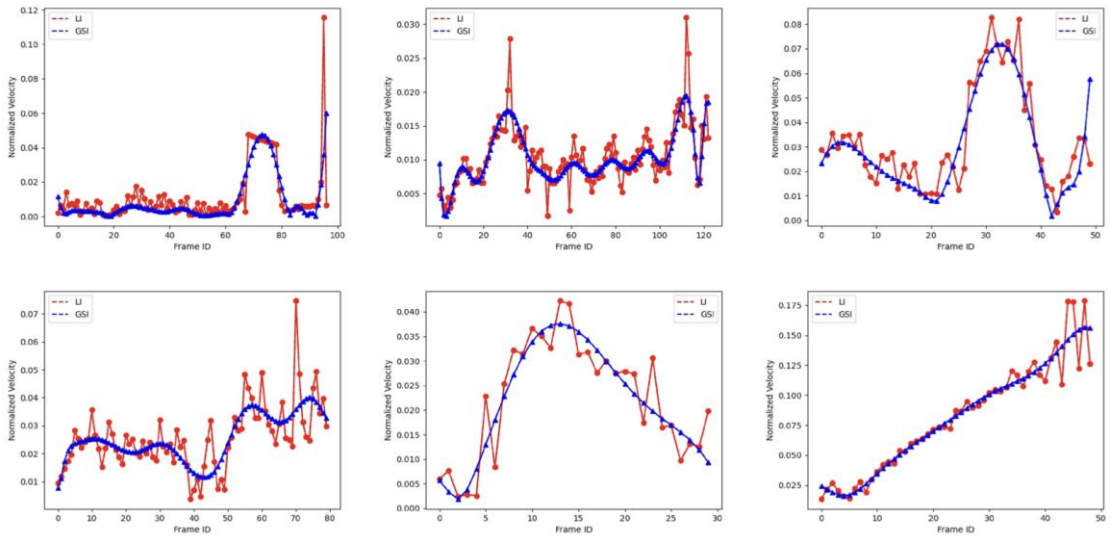


Fig. 5: Comparison of normalized velocity between the trajectories after applying linear interpolation (LI, in red) and Gaussian-smoothed interpolation (GSI, in blue). The x-coordinate represents the frame id, and the y-coordinate is the normalized velocity.

Điều này dẫn đến kết luận thú vị sau: Mặc dù Matching Cascade có thể làm giảm sự liên kết nhầm lẫn ở các trình theo dõi kém, thay vào đó, nó sẽ hạn chế hiệu suất của các trình theo dõi mạnh hơn.

Phân tích bổ sung cho GSI: *(Hình bên dưới)*

Việc ước tính tốc độ là cần thiết để phân tích hành động và lợi ích cho dự án Xây dựng hệ thống giao thông thông minh (ITSs - Intelligent Transport Systems). Nhóm tác giả đã so sánh vận tốc chuẩn hoá giữa các quỹ đạo sau khi áp dụng nội suy tuyến tính (LI) và phép nội suy làm mịn (GSI) trong hình dưới đây. Cụ thể, sáu quỹ đạo từ DeepSORT trên bộ xác thực MOT17 được sử dụng, toạ độ x và y biểu diễn frame id và normalized velocity tương ứng.

Qua biểu đồ có thể thấy rằng vận tốc của quỹ đạo với LI rung lắc dữ dội, chủ yếu do phát hiện nhiễu. Thay vào đó, quỹ đạo có GSI có vận tốc ổn định hơn, chính nó cũng đem đến cho nhóm tác giả một góc nhìn khác để hiểu GSI rằng GSI là một bộ lọc phát hiện tiếng ồn có thể hoạt động ổn định và chính xác hơn.

5.3 **Kết quả chính**:

Nhóm tác giả so sánh StrongSORT, StrongSORT+ (StrongSORT + AFLink) và StrongSORT++ (StrongSORT + AFLink + GSI) với các công cụ theo dõi tiên tiến nhất trên bộ thử nghiệm MOT17, MOT20, DanceTrack và KITTI, như được hiển thị trong Bảng IV, V, VI và VII tương ứng. Đáng chú ý, việc so sánh FPS một cách công bằng là rất khó vì tốc độ được yêu cầu bởi mỗi phương pháp phụ thuộc vào thiết bị nơi chúng được triển khai và thời

TABLE IV: Comparison with state-of-the-art MOT methods on the MOT17 test set. "*" represents our reproduced version. "(w/o LI)" means abandoning the offline linear interpolation procedure. The two best results for each metric are bolded and highlighted in red and blue.

TABLE VI: Comparison with state-of-the-art MOT methods on the DanceTrack test set. The two best results for each metric are bolded and highlighted in red and blue.

Method	Ref.	HOTA(↑)	IDF1(↑)	MOTA(↑)	AssA(↑)	DetA(↑)
CenterTrack [66]	ECCV2020	41.8	35.7	86.8	22.6	78.1

TABLE V: Comparison with state-of-the-art MOT methods on the MOT20 test set. "*" represents our reproduced version.

TABLE V: Comparison with state-of-the-art MOT methods on the MOT20 test set. "*" represents our reproduced version. led and "(w/o LI)" means abandoning the offline linear interpolation procedure. The two best results for each metric are bolded and highlighted in red and blue.

mode	Method	Ref.	HOTA(↑)	IDF1(↑)	MOTA(↑)	AssA(↑)	DetA(↑)	IDs(↓)	FPS(↑)	PS(↑)
online	SORT [3]	ICIP2016	36.1	45.1	42.7	35.9	36.7	4,470	**57.3**	**57.3**
	Tracktor++ [1]	ICCV2019	42.1	52.7	52.6	42.0	42.3	1,648	1.2	1.2
	CSTrack [28]	TIP2022	54.0	68.6	66.6	54.0	54.2	3,196	4.5	4.5
	FairMOT [64]	IJCV2021	54.6	67.3	61.8	54.7	54.7	5,243	13.2	13.2
	CrowdTrack [42]	AVSS2021	55.0	68.2	70.7	52.6	57.7	3,198	9.5	9.5
	RelationTrack [59]	TMM2022	56.5	70.5	67.2	56.4	56.8	4,243	4.3	4.3
	OC-SORT* (w/o LI) [7]	arxiv2022	60.5	74.4	**73.1**	**60.8**	**60.5**	**1,307**	-	-
	ByteTrack* (w/o LI) [63]	ECCV2022	**60.9**	**74.9**	**75.7**	59.9	**62.0**	1,347	**17.5**	**17.5**
	DeepSORT* [55]	ICIP2017	57.1	69.6	71.8	55.5	59.0	1,418	3.2	3.2
	StrongSORT	ours	**61.5**	**75.9**	72.2	**63.2**	59.9	**1,066**	1.5	1.5
offline	TBooster [49]	TMM2022	42.5	53.4	54.6	41.4	43.8	1,674	0.1	0.1
	MPNTrack [6]	CVPR2020	46.8	59.1	57.6	47.3	46.6	1,210	6.5	6.5
	MAATrack [43]	WACVw2022	57.3	71.2	73.9	55.1	59.7	1,331	**14.7**	**14.7**
	ReMOT [58]	IVC2021	61.2	73.1	**77.4**	58.7	**63.9**	1,789	0.4	0.4
	OC-SORT [7]	arxiv2022	**62.1**	75.9	75.5	-	-	**913**	-	-
	ByteTrack* [63]	ECCV2022	61.2	75.1	**76.5**	60.0	**62.6**	1,120	**17.5**	**17.5**
	StrongSORT+	ours	61.6	**76.3**	72.2	**63.6**	59.9	1,045	1.5	1.5
	StrongSORT++	ours	**62.6**	**77.0**	73.8	**64.0**	61.3	**770**	1.4	1.4

gian dành cho việc phát hiện thường bị loại trừ đối với các trình theo dõi theo dõi phát hiện.
MOT17: *(Kết quả tại bảng IV)* StrongSORT++ đứng đầu trên MOT17 và
mang lại sự liên kết chính xác và vượt trội hơn các trình theo dõi còn lại một khoảng cách lớn. Điều đáng lưu ý là phiên bản tái nâng cấp DeepSORT của nhóm tác giả cũng đã thực hiện tốt trên tiêu chuẩn.

MOT20: *(Kết quả tại bảng V)* Tập dữ liệu trong MOT20 được lấy từ nhiều nơi đông đúc người và các vật thể, tỷ lệ nhiễu cao hơn đồng nghĩa với việc nguy cơ bỏ sót các phát hiện và liên kết.
StrongSORT++ vẫn đứng đầu và đạt được thành tựu đáng kể hơn các trình theo dõi khác. Lưu ý rằng nhóm tác giả sử dụng chính xác các siêu tham số tương tự MOT17, ngụ ý khả năng khái quát hoá các phương pháp. Hiệu suất phát hiện của nó hơi kém so với một số trình theo dõi khác vì nhóm tác giả sử dụng cùng một ngưỡng điểm phát hiện như MOT17, dẫn đến nhiều phát hiện thiếu.

DANCETRACK:*(Kết quả tại bảng VI)* StrongSORT++ của nhóm tác giả cũng đạt được kết quả tốt nhất trên điểm chuẩn của DanceTrack cho hầu hết các số liệu. Bởi vì tập dữ liệu này ít

TABLE VI: Comparison with state-of-the-art MOT methods on the DanceTrack test set. The two best results for each metric are bolded and highlighted in red and blue.

Method	Ref.	HOTA(↑)	IDF1(↑)	MOTA(↑)	AssA(↑)	DetA(↑)
CenterTrack [66]	ECCV2020	41.8	35.7	86.8	22.6	78.1
FairMOT [64]	IJCV2021	39.7	40.8	82.2	23.8	66.7
TransTrack [45]	arxiv2020	45.5	45.2	88.4	27.5	75.9
TraDes [56]	CVPR2021	43.3	41.2	86.2	25.4	74.5
ByteTrack [63]	ECCV2022	47.7	53.9	**89.6**	32.1	71.0
MOTR [61]	ECCV2022	54.2	51.5	79.7	**40.2**	73.5
OC-SORT [7]	arxiv2022	**55.1**	**54.2**	89.4	38.0	**80.3**
StrongSORT++	ours	**55.6**	**55.2**	**91.1**	**38.6**	**80.7**

TABLE VII: Comparison with state-of-the-art MOT methods on the KITTI test set. The two best results for each metric are bolded and highlighted in red and blue.

		Car				Pedestrian			
Method	**Ref.**	**HOTA(↑)**	**MOTA(↑)**	**AssA(↑)**	**IDs(↓)**	**HOTA(↑)**	**MOTA(↑)**	**AssA(↑)**	**IDs(↓)**
AB3D [53]	IROS2020	69.99	83.61	69.33	**113**	37.81	38.13	44.33	**181**
MPNTrack [6]	CVPR2020	-	-	-	-	45.26	46.23	47.28	397
CenterTrack [66]	ECCV2020	73.02	88.83	71.20	254	40.35	53.84	36.93	425
QD-3DT [23]	TPAMI2022	72.77	85.94	72.19	**206**	41.08	51.77	38.82	717
QDTrack [34]	CVPR2021	68.45	84.93	65.49	313	41.12	55.55	38.10	487
LGMTracker [48]	ICCV2021	73.14	87.60	72.31	448	-	-	-	-
PermaTrack [46]	ICCV2021	**77.42**	**90.85**	**77.66**	275	47.43	65.05	43.66	483
OC-SORT [7]	arxiv2022	76.54	90.28	76.39	250	**54.69**	**65.14**	**59.08**	204
StrongSORT++	ours	**77.75**	**90.35**	**78.20**	440	**54.48**	**67.38**	**57.31**	**178**

tập trung vào các đặc điểm ngoại hình, nhóm tác giả đã bỏ qua các tối ưu hoá liên quan đến ngoại hình ở đây (Bot và EMA), tập trung nhiều vào các chuyển động và đạt được kết quả tốt hơn nhiều, chứng tỏ tính ưu việt của phương pháp.

KITTI:*(Kết quả tại bảng VII)* Trên tập dữ liệu KITTI, để đơn giản nhóm tác giả chỉ áp dụng hai thủ thuật (ECC, Kalman) và hai thuật toán được đề xuất (AFLink, GSI) cho nhiệm vụ lái

xe tự động. Kết quả cho thấy StrongSORT++ đạt được kết quả tốt cho ô tô và hiệu suất vượt trội với người đi bộ.

5.4 **Kết quả định tính**:

Hình dưới đây trực quan hoá một số kết quả của StrongSORT++ trên các tập dữ liệu MOT17, MOT20, DanceTrack và KITTI. Kết quả của MOT17 cho thấy tính hiệu quả của phương pháp trong các tình huống thông thường và hoạt động tốt khi camera đang chuyển động. Hơn nữa, kết quả của MOT20-04 cho thấy hiệu suất tuyệt vời của StrongSORT++ trong các tình huống tắc nghẽn nghiêm trọng. Kết quả của DanceTrack và KITTI chứng minh tính hiệu quả của StrongSORT++ trong khi phải đối mặt với các vấn đề về kiểu chuyển động phức tạp và tốc độ khung hình thấp.

5.4 **Hạn chế**:

StrongSORT và StrongSORT++ vẫn còn một số hạn chế. Một trong những quan tâm là tốc độ chạy tương đối chậm so với các theo dõi chung và một số theo dõi độc lập không phụ thuộc vào ngoại hình. Vấn đề này chủ yếu do mô hình DeepSORT, yêu cầu một bộ phát hiện và ngoại hình bổ sung, trong khi AFLink và GSI được đề xuất là hai thuật toán nhẹ. Bên cạnh đó, mặc dù phương pháp của nhóm tác giả hoạt động tốt trên các chỉ số IDF1 và HOTA, nhưng vẫn có MOTA thấp hơn một chút trên MOT17 và MOT20, chủ yếu do nhiều lần phát hiện bị thiếu do ngưỡng điểm phát hiện cao.

Đối với AFLink, mặc dù nó hoạt động tốt trong việc khôi phục các liên kết bị thiếu, nhưng nó không thể giải quyết được vấn đề gán sai liên kết. Cụ thể, AFLink không thể chia những quỹ đạo ID bị lẫn lộn thành các tracklet chính xác. Công việc tương lai cần phát triển các chiến lược kết nối toàn cầu mạnh mẽ và linh hoạt hơn.

6. Kết Luận

Trong bài viết này, nhóm tác giả đã xem lại công cụ theo dõi cổ điển DeepSORT và nâng cấp nó bằng các mô-đun mới cũng như một số thủ thuật suy luận. Công cụ theo dõi mới, StrongSORT, có thể đóng vai trò là baseline mạnh mẽ mới cho nhiệm vụ của MOT.

Họ cũng đề xuất hai thuật toán nhẹ và không có giao diện, AFLink và GSI, để giải quyết các vấn đề về liên kết bị thiếu và phát hiện thiếu. Các thử nghiệm cho thấy chúng có thể được áp dụng và mang lại lợi ích cho nhiều thiết bị theo dõi tiên tiến khác nhau với chi phí tính toán tăng thêm không đáng kể.

Bằng cách tích hợp StrongSORT với AFLink và GSI, công cụ theo dõi kết quả StrongSORT++ đạt được kết quả tiên tiến trên nhiều điểm chuẩn, tức là MOT17, MOT20, DanceTrack và KITTI.

Sự Nhìn Nhận

Dự án này được hỗ trợ bởi Quỹ khoa học tự nhiên quốc gia Trung Quốc (Chinese National Natural Science Foundation) dưới sự tài trợ (62076033, U1931202) và bằng tiến sĩ xuất sắc BUPT. Quỹ sinh viên (CX2022145).

Tài liệu tham khảo

[1] Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Track- ing without bells and whistles. In: Proceedings of the IEEE/CVF International Conference on Computer Vi- sion. pp. 941–951 (2019)

[2] Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. EURASIP Journal on Image and Video Processing 2008, 1–10 (2008)

[3] Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. IEEE (2016)

[4] Bochinski, E., Eiselein, V., Sikora, T.: High-speed tracking-by-detection without using image information. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS). pp. 1–6. IEEE (2017)

[5] Bochinski, E., Senst, T., Sikora, T.: Extending iou based multi-object tracking by visual information. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–6. IEEE (2018)

[6] Braso´, G., Leal-Taixe´, L.: Learning a neural solver for multiple object tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6247–6257 (2020)

[7] Cao, J., Weng, X., Khirodkar, R., Pang, J., Kitani, K.: Observation-centric sort: Rethinking sort for robust multi-object tracking. arXiv preprint arXiv:2203.14360 (2022)

[8] Dai, P., Wang, X., Zhang, W., Chen, J.: Instance segmen- tation enabled hybrid data association and discriminative hashing for online multi-object tracking. IEEE Transac- tions on Multimedia 21(7), 1709–1723 (2018)

[9] Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cre- mers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixe´, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)

[10] Du, Y., Tong, Z., Wan, J., Zhang, B., Zhao, Y.: Pami-ad: An activity detector exploiting part-attention and motion information in surveillance videos. In: 2022 IEEE Inter- national Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6. IEEE (2022)

[11] Du, Y., Wan, J., Zhao, Y., Zhang, B., Tong, Z., Dong, J.: Giaotracker: A comprehensive framework for mc-mot with global information and optimizing strategies in visdrone 2021. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2809– 2819 (2021)

[12] Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)

[13] Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient max-imization. IEEE transactions on pattern analysis and machine intelligence 30(10), 1858–1865 (2008)

[14] Ferna´ndez Llorca, D., Herna´ndez Mart´ınez, A., Garc´ıa Daza, I.: Vision-based vehicle speed estimation: A survey. IET Intelligent Transport Systems 15(8), 987–1005 (2021)

[15] Fu, Z., Angelini, F., Chambers, J., Naqvi, S.M.: Multi- level cooperative fusion of gm-phd filters for online mul- tiple human tracking. IEEE Transactions on Multimedia 21(9), 2277–2291 (2019)

[16] Gao, J., Nevatia, R.: Revisiting temporal model- ing for video-based person reid. arXiv preprint arXiv:1805.02104 (2018)

[17] Gao, T., Pan, H., Wang, Z., Gao, H.: A crf-based framework for tracklet inactivation in online multi-object tracking. IEEE Transactions on Multimedia 24, 995– 1007 (2021)

[18] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceed- ing yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)

[19] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32(11), 1231–1237 (2013)

[20] Han, S., Huang, P., Wang, H., Yu, E., Liu, D., Pan, X.: Mat: Motion-aware multi-object tracking. Neurocomput- ing (2022)

[21] He, J., Huang, Z., Wang, N., Zhang, Z.: Learnable graph matching: Incorporating graph partitioning with deep feature learning for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5299–5309 (2021)

[22] Hofmann, M., Haag, M., Rigoll, G.: Unified hierarchi- cal multi-object tracking using global data association. In: 2013 IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS). pp. 22–28. IEEE (2013)

[23] Hu, H.N., Yang, Y.H., Fischer, T., Darrell, T., Yu, F., Sun, M.: Monocular quasi-dense 3d object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[24] Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82D, 35–45 (1960)

[25] Khurana, T., Dave, A., Ramanan, D.: Detecting invisible people. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3174–3184 (2021)

[26] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2014)

[27] Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly 2(1-2), 83–97 (1955)

[28] Liang, C., Zhang, Z., Zhou, X., Li, B., Zhu, S., Hu, W.: Rethinking the competition between detection and reid in multiobject tracking. IEEE Transactions on Image Processing 31, 3182–3196 (2022)

Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1, pp. 666–673. IEEE (2006)

[37] Possegger, H., Mauthner, T., Roth, P.M., Bischof, H.: Occlusion geodesics for online multi-object tracking. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1306–1313 (2014)

[38] Redmon, J., Farhadi, A.: Yolov3: An incremental im- provement. arXiv preprint arXiv:1804.02767 (2018)

[39] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015)

[40] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European conference on computer vision. pp. 17–35. Springer (2016)

[41] Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., Sun, J.: Crowdhuman: A benchmark for detecting human in a crowd. arXiv preprint arXiv:1805.00123 (2018)

[42] Stadler, D., Beyerer, J.: On the performance of crowd- specific detectors in multi-pedestrian tracking. In: 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–12. IEEE (2021)

[43] Stadler, D., Beyerer, J.: Modelling ambiguous assign- ments for multi-person tracking in crowds. In: Proceedings of the IEEE/CVF Winter Conference on Applica- tions of Computer Vision. pp. 133–142 (2022)

[29] Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixe´, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision 129(2), 548–578 (2021)

[30] Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., Gu, J.: A strong baseline and batch normalization neck for deep person re-identification. IEEE Transactions on Multimedia 22(10), 2597–2609 (2019)

[31] Milan,A.,Leal-Taixe´,L.,Reid,I.,Roth,S.,Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)

[32] Naiel, M.A., Ahmad, M.O., Swamy, M., Lim, J., Yang, M.H.: Online multi-object tracking via robust collabora- tive model and sample selection. Computer Vision and Image Understanding 154, 94–107 (2017)

[33] Pang, B., Li, Y., Zhang, Y., Li, M., Lu, C.: Tubetk: Adopting tubes to track multi-object in a one-step training model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6308–6318 (2020)

[34] Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., Yu, F.: Quasi-dense similarity learning for multiple object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 164–173 (2021)

[35] Peng, J., Wang, T., Lin, W., Wang, J., See, J., Wen, S., Ding, E.: Tpm: Multiple object tracking with tracklet-plane matching. Pattern Recognition 107, 107480 (2020)

[36] Perera, A.A., Srinivas, C., Hoogs, A., Brooksby, G., Hu, W.: Multi-object tracking through simultaneous long occlusions and split-merge conditions. In: 2006 IEEE

[44] Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20993–21002 (2022)

[45] Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)

[46] Tokmakov, P., Li, J., Burgard, W., Gaidon, A.: Learning to track with object permanence. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10860–10869 (2021)

[47] Wang, B., Wang, G., Chan, K.L., Wang, L.: Tracklet association by online target-specific metric learning and coherent dynamics estimation. IEEE transactions on pat- tern analysis and machine intelligence 39(3), 589–602 (2016)

[48] Wang, G., Gu, R., Liu, Z., Hu, W., Song, M., Hwang, J.N.: Track without appearance: Learn box and tracklet embedding with local and global motion patterns for vehicle tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9876–9886 (2021)

[49] Wang, G., Wang, Y., Gu, R., Hu, W., Hwang, J.N.: Split and connect: A universal tracklet booster for multi-object tracking. IEEE Transactions on Multimedia pp. 1– 1 (2022). https://doi.org/10.1109/TMM.2022.3140919

[50] Wang, G., Wang, Y., Zhang, H., Gu, R., Hwang, J.N.: Ex- ploit the connectivity: Multi-object tracking with track- letnet. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 482–490 (2019)

puter Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 1–21. Springer (2022)

[64] Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Com- puter Vision 129(11), 3069–3087 (2021)

[51] Wang, Q., Zheng, Y., Pan, P., Xu, Y.: Multiple object

tracking with correlation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3876–3886 (2021)

[52] Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: European Conference on Computer Vision. pp. 107–122. Springer (2020)

[53] Weng, X., Wang, J., Held, D., Kitani, K.: 3d multi-object tracking: A baseline and new evaluation metrics. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10359–10366. IEEE (2020)

[54] Williams, C., Rasmussen, C.: Gaussian processes for regression. Advances in neural information processing systems 8 (1995)

[55] Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)

[56] Wu, J., Cao, J., Song, L., Wang, Y., Yang, M., Yuan, J.: Track to detect and segment: An online multi-object tracker. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12352–12361 (2021)

[57] Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixe´, L., Alameda-Pineda, X.: How to train your deep multi-object tracker. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6787– 6796 (2020)

[58] Yang, F., Chang, X., Sakti, S., Wu, Y., Nakamura, S.: Remot: A model-agnostic refinement for multiple object tracking. Image and Vision Computing 106, 104091 (2021)

[59] Yu, E., Li, Z., Han, S., Wang, H.: Relationtrack: Relation- aware multiple object tracking with decoupled represen- tation. IEEE Transactions on Multimedia pp. 1–1 (2022). https://doi.org/10.1109/TMM.2022.3150169

[60] Yu, F., Li, W., Li, Q., Liu, Y., Shi, X., Yan, J.: Poi: Multiple object tracking with high performance detection and appearance feature. In: European Conference on Computer Vision. pp. 36–42. Springer (2016)

[61] Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII. pp. 659–675. Springer (2022)

[62] Zhang, S., Benenson, R., Schiele, B.: Citypersons: A diverse dataset for pedestrian detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3221 (2017)

[63] Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII. pp. 1–21. Springer (2022)

[64] Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Com- puter Vision 129(11), 3069–3087 (2021)

[65] Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: European conference on computer vision. pp. 868–884. Springer (2016)

[66] Zhou, X., Koltun, V., Kra ̈henbu ̈hl, P.: Tracking objects as points. In: European Conference on Computer Vision. pp. 474–490. Springer (2020)

[67] Zhu, Y., Zhou, K., Wang, M., Zhao, Y., Zhao, Z.: A comprehensive solution for detecting events in complex surveillance videos. Multimedia Tools and Applications 78(1), 817–838 (2019)