NGHIÊN CỨU CÁC PHƯƠNG PHÁP THEO DÕI ĐỐI TƯỢNG SỬ DỤNG KIẾN TRÚC MẠNG SIAMESE

Hoàng Đình Thắng[1*], Trần Quốc Long[2], Thái Trung Kiên[1]

Tóm tắt: *Theo dõi đối tượng đã trở thành một trong những chủ đề nghiên cứu hấp dẫn nhất trong thị giác máy tính và ngày càng phát triển trong ứng dụng thương mại cũng như nghiên cứu học thuật. Công nghệ học sâu đã thực sự nâng tầm bài toán theo dõi đối tượng, bằng cách cung cấp khả năng học các đặc trưng trừu tượng giúp bám bắt đối tượng mạnh mẽ hơn. Các mô hình theo dõi dựa trên Siamese đã đạt được hiệu suất rất cao trong việc theo dõi đối tượng trực quan. Bài báo này trình bày phân tích sâu về khả năng học và suy luận của tám mô hình theo dõi dựa trên mạng nơron Siamese trong những năm gần đây. Kết quả định lượng từ tám bài báo được so sánh trên các bộ đánh giá theo dõi trực quan phổ biến OTB2015, UAV123, VOT2018 và LaSOT.*

Từ khóa: Thị giác máy tính; Theo dõi đối tượng; Học sâu; Mạng nơron Siamese.

1. ĐẶT VẤN ĐỀ

Theo dõi đối tượng trực quan nhằm theo dõi đối tượng mục tiêu trong mỗi khung hình của một chuỗi video. Đây là một lĩnh vực nghiên cứu mở và hấp dẫn với nhiều ứng dụng thực tế, bao gồm rôbốt tự hành [1], giám sát [2], ô tô tự lái [3], theo dõi nhìn từ trên không [4]. Từ những đột phá về học sâu (DL) trong cuộc thi nhận dạng hình ảnh quy mô lớn ImageNet (ILSVRC) và cả thử thách theo dõi đối tượng trực quan (VOT), các phương pháp dựa trên DL đã thu hút sự quan tâm của cộng đồng theo dõi trực quan để cung cấp các trình theo dõi mạnh mẽ. Mặc dù ban đầu mạng nơron tích chập (CNN) là mạng thống trị, nhưng một loạt các kiến trúc như mạng nơron hồi quy (RNN), mã tự động (AE), mạng sáng tạo đối nghịch (GAN) và đặc biệt là mạng nơron Siamese (SNN) & mạng nơron tùy biến hiện đang được quan tâm nghiên cứu. SNN được sử dụng rộng rãi cho các trình theo dõi trực quan trong vài năm qua. Với các cặp mục tiêu và vùng tìm kiếm, các mạng SNN hai luồng tính toán cùng một hàm để tạo ra một bản đồ tương tự. SNN chủ yếu hướng tới việc khắc phục những hạn chế của các CNN sâu được đào tạo trước và tận dụng tối đa việc học từ đầu đến cuối cho các ứng dụng thời gian thực.

Gần đây, có một số bài báo [5-8] xem xét các trình theo dõi trực quan dựa trên học sâu. Ví dụ, [5, 6] phân loại các phương pháp thủ công và học sâu thành trình theo dõi lọc tương quan và lọc không tương quan. Bài báo [7] phân loại trình theo dõi dựa trên học sâu theo cấu trúc, chức năng và đào tạo. [8] trình bày các hướng nghiên cứu mới nhất về theo dõi chuyển động dựa trên mạng nơron tích chập. Công trình [9] đặc biệt nghiên cứu một số trình theo dõi sử dụng mạng Siamese dựa trên các khía cạnh nhánh của mạng, các tầng và đào tạo mạng. Tuy nhiên, bài báo này không bao gồm những trình theo dõi mới nhất hiện nay.

Mục tiêu chính của bài báo này là phân tích sâu các trình theo dõi dựa trên kiến trúc Siamese tiên tiến nhất từ năm 2018 trở lại đây, cụ thể là: SiamRPN [19], DaSiamRPN [20], SiamDW [21], SiamRPN++ [22], SiamAttn [23], SiamBAN [24], SiamCAR [25], SiamRCNN [26]. Ngoài ra, bài báo này còn nêu ra các thuộc tính trực quan thách thức nhất cho tập dữ liệu và tập dữ liệu OTB2015, và UAV123, VOT2018 và LaSOT. Bằng cách làm như vậy, những thách thức chính nhất của mỗi tập dữ liệu đối với các trình theo dõi gần đây được chỉ ra.

Đóng góp chính được tóm tắt như sau: Bài báo giới thiệu một số trình theo dõi trực quan dựa trên kiến trúc Siamese tiểu biểu từ năm 2018 đến năm 2020 được đăng trên các hội thảo hàng đầu về thị giác máy tính.

Bài báo này được trình bày theo thứ tự như sau: Phần 2 trình bày những thách thức và phương pháp đánh giá mô hình theo dõi chuyển động; Các phương pháp theo dõi chuyển động sử dụng mạng Siamese những năm gần đây, so sánh thực nghiệm của các phương pháp được thực hiện trong phần 3; Cuối cùng, phần 4 trình bày tóm tắt các kết luận.

2. THÁCH THỨC VÀ PHƯƠNG PHÁP ĐÁNH GIÁ THEO DÕI CHUYỂN ĐỘNG

2.1. Các chỉ số đánh giá phương pháp theo dõi đối tượng đơn

(i) Tính chính xác (Accuracy): Điểm chồng chéo được tính theo $S = \frac{|b_t \cap b_g|}{|b_t \cup b_g|}$. Trong đó, b_g, b_t, $\cap$, $\cup$ và $|.|$ lần lượt là hộp bao quanh đối tượng (BB ground-truth), hộp bao quanh được ước lượng bởi trình theo dõi, toán tử giao nhau, toán tử liên hợp và số lượng pixel trong vùng kết quả. Điểm chồng chéo cho biết thành công của trình theo dõi hình ảnh trong một khung hình bằng cách so sánh với một ngưỡng. Tính chính xác được tính bằng điểm số chồng chéo trung bình (AOS) trong quá trình theo dõi khi các ước tính của trình theo dõi trực quan chồng chéo với ground-truth.

(ii) Điểm độ bền (Robustness)/thất bại (failure): Điểm độ bền hoặc thất bại được xác định là số lần khởi động lại bắt buộc khi trình theo dõi mất hoặc trôi dạt mục tiêu trong quá trình theo dõi. Thất bại được xác định khi điểm chồng chéo giảm về không.

(iii) Độ chồng chéo trung bình dự kiến (EAO): Điểm số này là sự kết hợp của điểm số tính chính xác và độ bền. Với chuỗi video gồm N_s khung hình, điểm EAO được tính bằng $\widehat{\emptyset}_{N_s} = \langle \frac{1}{N_s} \sum_{i=1}^{N_s} \emptyset_i \rangle$. Trong đó, $\emptyset_i$ được xác định là giá trị trung bình chồng chéo mỗi khung hình cho đến khi kết thúc chuỗi khung hình.

(iv) Diện tích dưới đường cong (AUC): Điểm AUC được xác định tỷ lệ thành công trung bình (được chuẩn hóa từ 0 đến 1) theo các ngưỡng trước khi xác định. Để xếp hạng các phương pháp theo dõi trực quan dựa trên hiệu suất tổng thể, điểm AUC tổng AOS của các phương pháp theo dõi trực quan trên mỗi chuỗi video.

(v) Lỗi vị trí trung tâm (CLE)/Độ chính xác chuẩn hóa: CLE hay chỉ số độ chính xác là khoảng cách Euclide trung bình giữa các vị trí ground-truth của mục tiêu với các vị trí ước tính bởi trình theo dõi trực quan. Độ chính xác chuẩn hóa nhằm mục đích làm giảm độ nhạy của CLE đối với kích thước của BB và độ phân giải khung hình.

(vi) Biểu đồ chính xác (Precision plot): Với CLE cho mỗi ngưỡng khác nhau, biểu đồ chính xác hiển thị phần trăm khung hình video trong đó các vị trí ước tính cách vị trí ground-truth một ngưỡng xác định.

(vii) Biểu đồ thành công (Success plot): Với độ chính xác khác nhau được tính toán trên mỗi ngưỡng, biểu đồ thành công đo lường tỷ lệ phần trăm các khung hình trong đó độ chồng chéo khung hình ước tính với khung hình ground-truth lớn hơn một ngưỡng nhất định.

2.2. Thách thức trên các bộ dữ liệu đánh giá

Bộ dữ liệu đánh giá theo dõi chuyển động trực quan đã được đưa ra để cung cấp các đánh giá công bằng và chuẩn về các thuật toán theo dõi đối tượng đơn. Những bộ dữ liệu này được phân loại dựa trên ứng dụng theo dõi đối tượng chung chung và theo dõi từ trên không. Các tập dữ liệu này chứa các chuỗi, khung hình, thuộc tính và lớp (hoặc cụm) khác nhau. Các thuộc tính bao gồm thay đổi độ sáng (IV), thay đổi tỷ lệ (SV), tắc (OCC), biến dạng (DEF), chuyển động mờ (MB), chuyển động nhanh (FM), quay trong mặt phẳng (IPR), xoay ngoài mặt phẳng (OPR), ngoài tầm nhìn (OV), nền lộn xộn (BC), độ phân giải thấp (LR), thay đổi tỷ lệ khung hình (ARC), chuyển động camera (CM), tắc hoàn toàn (FOC), tắc một phần (POC), đối tượng tương tự (SIB), thay đổi điểm nhìn (VC), độ tương phản thấp (LC), camera phóng to (ZC), thời lượng dài (LD), ánh sáng mờ (DL), rung máy (CS), xoay (ROT), thay đổi nền nhanh (FBC), thay đổi chuyển động (MOC), thay đổi màu đối tượng (OCO), độ phức tạp của cảnh (SCO), chuyển động tuyệt đối (AM), chuyển động camera nhanh (FCM), và theo dõi dài hạn (LT).

OTB2015 [13] là bộ dữ liệu đáng giá theo dõi đối tượng chung chung, được phát triển từ OTB2013 với mục đích so sánh hiệu suất không thiên lệch. OTB2015 chứa 100 video được thu

thập từ các chuỗi theo dõi thường được sử dụng. Đánh giá dựa trên hai chỉ số: biểu đồ chính xác và biểu đồ thành công.

UAV123 [14]. Theo dõi từ các góc nhìn từ trên không đã được phát triển trong những năm gần đây và có nhiều ứng dụng được nghiên cứu. UAV123 chứa các chuỗi video HD thực tế cung cấp tập dữ liệu theo dõi theo chế độ xem trên không thưa thớt và độ cao thấp. Đánh giá dựa trên hai chỉ số: biểu đồ chính xác và biểu đồ thành công.

VOT2018 [15] nhằm cung cấp một tập dữ liệu nhỏ, đa dạng và đầy đủ từ tập dữ liệu hiện có, được gán nhãn bởi các hình chữ nhật có thể xoay bao quanh đối tượng (BB) và các thuộc tính trực quan. Chỉ số đánh giá gồm: Độ chính xác (Accuracy), độ mãnh mẽ (Robustness), và độ chồng chéo trung bình dự kiến (EAO).

LaSOT [16] là bộ dữ liệu theo dõi đối tượng đơn lẻ quy mô lớn, đã được phát triển để giải quyết các vấn đề của bộ dữ liệu hiện có, chẳng hạn như quy mô nhỏ, thiếu chú thích chất lượng cao, dày đặc, chuỗi video ngắn và sai lệch danh mục. Các danh mục đối tượng là từ ImageNet và một số ứng dụng theo dõi trực quan với số lượng video trên mỗi danh mục bằng nhau. Tập hợp huấn luyện và thử nghiệm bao gồm chuỗi 1120 (2,3M khung hình) video và 280 (690K khung hình) video.

Bảng 1 so sánh các bộ dữ liệu đánh giá OBT2015, UAV123, VOT2018 và LaSOT trên các tiêu chí số lượng video (NoV), số lượng khung hình (NoF), số lớp (NoC), độ dài trung bình (AD), số lượng thuộc tính (NoA), thuộc tính.

Bảng 1. *So sánh các bộ dữ liệu đánh giá theo dõi chuyển động.*

Năm	Bộ dữ liệu	NoV	NoF	NoC	AD	NoA	Thuộc tính
2015	**OBT2015**	100	59K	16	19,8s	11	IV, SV, OCC, DEF, MB, FM, IPR, OPR, OV, BC, LR
2016	**UAV123**	123	113K	9	30,6s	12	IV, SV, FM, OV, BC, LR, ARC, CM, FCM, FOC, POC, SIB, VC
2018	**VOT2018**	60	21K	24	12s	12	IV, SV, OCC, DEF, MB, BC, ARC, CM, MOC, OCO, SCO, AM
2019	**LaSOT**	1400	3,5M	70	84,3s	14	IV, SV, DEF, MB, FM, OV, BC, LR, ARC, CM, FOC, POC, VC, ROT

3. CÁC PHƯƠNG PHÁP THEO DÕI ĐỐI TƯỢNG TRỰC QUAN SỬ DỤNG KIẾN TRÚC MẠNG SIAMESE

3.1. Tổng quan

Nhiệm vụ theo dõi đối tượng trực quan có thể được xây dựng như một bài toán học có nhiệm vụ chính là tìm vị trí mục tiêu tối ưu bằng cách cực tiểu hàm mất mát:

$$L(w) = \sum_{j=1}^{m} \gamma_j r\left(f\left(x_{j;}w\right), y_j\right) + \sum_k \lambda_k ||w_k||^2, \quad (1)$$

Trong đó, $r(f, y)$ tính toán phần dư tại mỗi vị trí không gian và $y_j \in R^{W \times H}$ là nhãn được chú thích. Tác động của r được thiết lập bởi γ_j và chuẩn hóa của w_k được thiết lập bởi λ_k.

Thuật toán theo dõi dựa trên mạng Siamese lấy (1) như nhiệm vụ khớp mẫu bằng cách xây dựng công thức như một vấn đề tương quan chéo (cross-correlation) f và học không gian nhúng $\varphi(\cdot)$ để tính toán vị trí nào trong vùng tìm kiếm có thể khớp nhất với mẫu.

$$f(x, z) = \varphi(x) * \varphi(z) + b * 1 \quad (2)$$

Trong đó, một nhánh học cách biểu diễn đặc trưng của mục tiêu z, nhánh còn lại học vùng tìm kiếm x.

SNN được sử dụng rộng rãi cho các trình theo dõi trực quan trong vài năm qua. Mục sau sẽ trình bày một số trình theo dõi điển hình từ năm 2018 đến 2020, được đăng ở các hội nghị hàng đầu về thị giác máy tính CVPR, ECCV.

3.2. Phương pháp sử dụng neo (anchor)

SiamRPN [19] đề xuất phương pháp gồm một mạng con Siamese để trích xuất đặc trưng và mạng con đề xuất vùng để sinh vùng chứa ứng cử viên. Trong mạng Siamese, hai nhánh sử dụng AlexNet [17] gồm chuỗi 5 tích chập *conv* kết hợp với *maxpool, BatchNorm2d* và *relu* để trích chọn đặc trưng từ ảnh mẫu và vùng tìm kiếm. Cặp đặc trưng này được đưa vào mạng sinh ứng cử viên gồm hai nhánh, một nhánh để phân lớp đối tượng với nền, và nhánh còn lại để sàng lọc ứng cử viên. Trong mỗi nhánh, đặc trưng từ mạng siamese qua conv sau đó kết hợp lại bởi phép tương quan chéo để cho bản đồ đặc trưng là 2k kênh hay 4k kênh ứng với nhánh phân lớp và hồi quy. Trong quá trình đào tạo, cặp ảnh mẫu được lấy ngẫu nhiên từ ImageNet [11] và liên tục từ Youtube-BB [12]. Mẫu đối tượng và mẫu để tìm được trích xuất từ hai khung hình của cùng một video. Đầu ra của mô hình là bản đồ tin cậy (CM)-thể hiện khu vực chứa đối tượng cần theo dõi. Quá trình đào đạo hàm mất mát cho phân lớp là hàm entropy chéo, và hàm L1 trơn với tọa độ được chuẩn hóa để hồi quy vị trí của đối tượng. SiameseRPN đào tạo từ đầu đến cuối dùng dùng Stochastic Gradient Descent (SGD), mạng con Siamese dùng backbone AlexNet đã được đào tạo trước trên Imagenet. Một số phép tăng cường dữ liệu được trong đó có phép biến đổi afin. Tỷ lệ neo được dùng là [0.33, 0.5, 1, 2, 3].

DaSiamRPN [20] sử dụng mạng AlexNet được đào tạo trước trên ImageNet như trong SiamRPN [19] làm hàm φ, trong đó tham số của 3 lớp tích chập *conv* đầu tiên được giữa nguyên, chỉ điều chỉnh 2 lớp tích chập cuối. Dữ liệu huấn luyện được lấy từ các bộ dữ liệu ImageNet-VID [11], Youtube-BB, ImageNet-DET [11] và MSCOCO [10]. DaSiamRPN đề xuất các chiến lược để cải thiện khái quát các đặc trưng đã học và loại bỏ sự phân bố không cân đối của dữ liệu. Để xử lý các video xám trong các bộ dữ liệu đánh giá, 25% cặp ảnh được chuyển sang chế độ xám trong quá trình đào tạo. Trong quá trình suy luận, phương pháp đề xuất sử dụng hàm mục tiêu mới phân biệt đối tượng gây nhiễu để sắp xếp lại ứng cử viên giống với mẫu cần tìm. Mô hình SiamRPN có khuynh hướng tìm đối tượng bất kỳ khi đối tượng ra khỏi khung nhìn hoặc tắc hoàn toàn, do đó dẫn đến trôi dạt theo dõi. Trong DaSiamRPN, khi gặp thất bại, chiến thuật tìm từ cục bộ đến toàn cục được sử dụng bằng việc tăng kích thước vùng tìm kiếm từ 255 lên 767.

SiamRPN++ [21] sử dụng ResNet[18] được đào tạo trước trên ImageNet làm backbone trích chọn đặc trưng bằng việc thay đổi một số tham số của mạng ResNet để khai thác tính năng sâu mà không làm mất đặc trưng dịch chuyển bất biến của mạng Siamese. Cụ thể, thay đổi stride và tăng dilate cho khối *conv4* và *conv5*, đồng thời 3 lớp tích chập 1x1 được gắn vào *conv3, conv4, conv5* của ResNet để giảm chiều đặc trưng về 256. SiamRPN++ đề xuất sử dụng tương quan chéo sâu-rộng (DW-Xcorr) để kết hợp thông tin giữa hai nhánh hiệu quả hơn so với mạng SiamRPN (số tham số ít hơn 10 lần so với SiamRPN). Mạng được đào tạo trên MSCOCO, ImageNet-DET, ImageNet-VID và Youtube-BB như DaSiamRPN. Cả đào tạo và suy luận đều sử dụng ảnh mẫu 127 pixel và khu vực tìm kiếm 255 pixel. SiamRPN++ đào tạo dùng SGD, hệ số học là 0.001 đối với 5 epoche đầu tiên, 15 epoche tiếp theo hệ số học giảm dần từ 0.005 đến 0.0005. Hàm mất mát tương tự như SiamRPN. Đầu ra của mô hình là điểm số và đường bao quang đối tượng.

SiamDW [22]. Việc nới rộng (padding) mạng dẫn đến sai lệch vị trí trong mô hình Siamese, SiamDW đề xuất dùng đơn vị CIR (Cropping-Inside Residual) để bỏ nới rộng khi xây dựng mạng Siamese dùng backbone ResNet, ResNeXt và Inception. SiamDW xây dựng mạng backbone sâu và rộng hơn bằng cách xếp chồng các đơn vị CIR. Tham số mô hình được khởi tạo với trọng số được đào tạo trước trên ImageNet, trong quá trình đào tạo đóng băng lớp tích chập đầu tiên, và tinh chỉnh các lớp khác. Đầu ra của mô hình là bản đồ tin cậy và bản đồ đặc tính. Bộ dữ liệu huấn luyện là ImageNet, ILSVRC-VID và Youtube-BB.

SiamAttn [23] bao gồm một môđun chú ý Siamese biến dạng (DSA), một mạng đề xuất vùng Siamese (SiamRPN) và một môđun tinh chỉnh vùng. Đặc trưng của 3 giai đoạn cuối được trích xuất và điều chế bởi môđun DSA. DSA sau đó tạo ra hai luồng đặc trưng chú ý và đưa vào các khối SiamRPN để dự đoán vùng theo dõi đơn. Môđun tinh chỉnh vùng cải thiện độ chính xác vị trí của vùng mục tiêu dự đoán. Kết quả theo dõi là hình bao và phân vùng đối tượng

SiamRCNN [26] được xây dựng từ FasterRCNN với khung sườn trích chọn đặc trưng là ResNet-101-FPN, tiếp theo là môđun đề xuất ứng cử viên, hai phần xác định lại đối tượng từ khung hình đầu tiên và khung hình trước đó, cuối cùng là thuật toán động xác định kết quả theo dõi là hình bao và phân vùng đối tượng. Mô hình được đào tạo đồng thời trên các bộ dữ liệu ImageNet-VID, YouTube-VOS 2018, GOT-10k và LaSOT. Phương pháp này dùng 18k video và 119k ảnh từ MSCOCO, so với SiamRPN++ dùng 384k video và 1867k ảnh tĩnh. Trong quá trình đào tạo, mô hình dùng chuyển động mờ, tỷ lệ xám, gamma, lật và tăng tỷ lệ.

3.3. Phương pháp không sử dụng neo

Các phương pháp dùng neo đề cập ở trên thường phải định nghĩa trước tham số neo để xác định chính xác quy mô và tỷ lệ của mục tiêu. Để thiết kế các tham số này cần phải có kinh nghiệm điều chỉnh và liên quan đến nhiều thủ thuật để đạt được hiệu suất tốt. Để không phải thiết lập các tham số neo, một số phương pháp không sử dụng neo đã được đề xuất.

SiamBAN [24] sử dụng ResNet-50 làm khung trích chọn đặc trưng tương tự như SiamRPN++, bằng việc thiết lập số trượt bằng 2 ở hai khối *conv4* và *conv5*, tỷ lệ atrous ở *conv4* và *conv5* lần lượt là 2 và 4. Đặc trưng thu được tại 3 khối conv3, conv4, conv5 ở mỗi nhánh sẽ được điều chỉnh về cùng số kênh 256 thông qua *conv*, sau đó được kết hợp lại với nhau thông qua DW-Xcorr. Hàm mất mát dùng trong mô hình gồm hàm phân lớp là hàm entropy chéo và hàm hồi quy là hàm IoU. Đầu ra của mô hình là đường bao quanh đối tượng và điểm số đối tượng. SiamBAN được đào tạo trên các bộ dữ liệu ImageNet-VID, YouTube-BB, MSCOCO, ImageNet-DET, GOT-10k và LaSOT dùng thuật toán tối ưu SGD.

SiamCAR [25] gồm hai mạng con, mạng con Siamese sử dụng ResNet-50 để trích chọn đặc trưng của mẫu và vùng tìm kiếm, đặc trưng thu được được điều chỉnh về cùng số kênh và kết nối thông tin với nhau thông qua DW-Xcorr để được bản đồ phản hồi đặc trưng R^* làm đầu vào cho mạng con phân lớp-hồi quy. Nhánh phân lớp và nhánh hồi quy của mạng con này gồm 5 tích chập *conv* cho phép xác định đối tượng cần tìm kiếm từ R^*. Hàm mất mát được dùng là hàm entropy chéo, IoU và hàm mất mát trung tâm. Bộ dữ liệu dùng để đào tạo gồm MSCOCO, ImageNet-DET, ImageNet-VID, YouTube-BB. Qua trình suy luận, mô hình dự đoán hộp bao quanh đối tượng trong khung hình hiện tại.

Tổng hợp và so sánh các phương pháp trên theo các tiêu chí khung sườn, dữ liệu huấn luyện, đặc tính khai thác, nền tảng ngôn ngữ, tốc độ chạy, và kết quả đầu ra được trình bày trong bảng 2.

Bảng 2. *So sánh kiến trúc các phương pháp theo dõi chuyển động sử dụng kiến trúc Siamese: DAF (đặc tính hình dạng sâu), CM (bản đồ tin cậy), OS (điểm đối tượng), BB (hộp bao đối tượng), FM (bản đồ đặc tính), SGM (mặt nạ phân vùng).*

Phương pháp	Backbone	Dữ liệu huấn luyện	Đặc tính	Ngôn ngữ	Nền tảng	Tốc độ	Đầu ra
SiamRPN	AlexNet	ImageNet, Youtube-BB	DAF	Python	PyTorch	160	CM
DaSiamRPN	AlexNet	ImageNet, Youtube-BB, ImageNet-DET, MSCOCO-DET	DAF	Python	PyTorch	160	CM
SiamRPN++	ResNet-50	ImageNet,	DAF	Python	PyTorch	35	OS,

		MSCOCO, ImageNet-DET, ImageNet-VID, Youtube-BB					BB
SiamDW	ResNet, ResNeXt, Inception	ImageNet, ImageNet-VID, Youtube-BB	DAF	Python	PyTorch	13.93	CM, FM
SiamAttn	ResNet-50	MSCOCO, YouTube-VOS, LaSOT, TrackingNet	DAF	Python	PyTorch	33	BB, SGM
Siam R-CNN	Faster R-CNN, ResNet-101-FPN	ImageNet-VID, MSCOCO, YouTube-VOS, GOT-10k, LaSOT	DAF	Python	TensorFlow	4.7	BB, SGM
SiamBAN	ResNet-50	ImageNet -VID, Youtube-BB, MSCOCO, ImageNet-DET, GOT-10k, LaSOT	DAF	Python	PyTorch	40	OS, BB
SiamCAR	ResNet-50	MSCOCO, ImageNet-DET, ImageNet-VID, Youtube-BB	DAF	Python	PyTorch	40	OS, BB

3.4. So sánh định lượng các phương pháp

Mục này so sánh các trình theo dõi dựa trên các độ đo định lượng đã đề cập ở mục 2.1 trên các bộ dữ liệu đánh giá OBT2015, UAV123, VOT2018 và LASOT. Kết quả được tổng hợp như bảng 3.

3.4.1. So sánh trên bộ dữ liệu OTB2015

Hai chỉ số được dùng để đánh giá trên bộ dữ liệu OTB2015 là AUC và biểu đồ chính xác (Prep.). Kết quả được lấy từ thực nghiệm của bài báo gốc cho thấy trình theo dõi SiamAttn đạt AUC và Prep. cao nhất lần lượt là 0.712 và 0.926. Các trình theo dõi SiamRPN++ và SiamBAN đạt chỉ số Prep. khá cao ứng với 0.923 và 0.910. SiamCAR không trình bày kết quả thử nghiệm trên bộ dữ liệu này.

3.4.2. So sánh trên bộ dữ liệu UAV123

Đối với bộ dữ liêu theo dõi từ trên không UAV123, SiamRPN và SiamDW không có kết quả thử nghiệm được trình bày trong bài báo. Hai chỉ số được dùng để đánh giá là biểu đồ thành công (Succ.) và biểu đồ chính xác (Prep.). Trình theo dõi SiamAttn có chỉ số thực nghiệm cao nhất trên cả hai chỉ số Succ.(0.650) và Prep.(0.845). Tiếp đến là SiamRCNN với Succ.(0.649)/Prep.(0.834). Các trình theo dõi đề xuất trước đó và trình theo dõi không dùng neo có chỉ số thấp hơn (xem bảng 3).

3.4.3. So sánh trên bộ dữ liệu VOT2018

Các trình theo dõi SiamRPN và DaSiamRPN đánh giá trên bộ VOT cũ hơn VOT2018. SiamAtnn tiếp tục cho thấy kết quả tốt nhất trong số các trình theo dõi được đề cập trong bài báo này với các chỉ số đánh giá A(0.63)/R(0.16)/EAO(0.47). Trình theo dõi SiamBAN cho thấy tính mãnh mẽ với chỉ số R thấp bằng 0.178 và chỉ số EAO (0.452) cao hơn các trình theo dõi khác.

3.4.4. So sánh trên bộ dữ liệu LaSOT

LaSOT là bộ liệu theo dõi đối tượng có thời lượng video dài. SiamRCNN với khả năng phát hiện lại đối tượng cho kết quả tốt nhất, với Succ./Prep. là 0.648/0.722. Tiếp đến là SiamAttn với Succ., Prep. lần lượt là 0.560 và 0.648. SiamBAN có kết quả cao thứ 3 với 0.514/0.598.

Từ kết quả thực nghiệm thấy rằng, trình theo dõi SiamAttn đạt kết quả tốt nhất trên các bộ dữ liệu OTB2015, UAV123 và VOT2018, còn SiamRCNN đạt kết qua tốt nhất trên bộ dữ liệu LaSOT. Trình theo dõi dùng neo tuy chạy chậm hơn nhưng kết quả thường cao hơn trình theo dõi không dùng neo.

Bảng 3. *So sánh kết quả các phương pháp theo dõi chuyển động sử dụng kiến trúc Siamese.*

Hội thảo	**Trình theo dõi**	**OTB2015**	**UAV123**	**VOT2018**	**LASOT**
		AUC/Prep.	**Succ./Prep.**	**A/R/EAO**	**Succ./Prep.**
CVPR2018	**SiamRPN**	0.637/0.851	-	-	-
ECCV2018	**DaSiamRPN**	-/0.88	0.586/0.796	-	-
CVPR2019	**SiamRPN++**	0.696/0.923	0.613/0.807	0.600/0.234/0.414	0.496/0.569
	SiamDW	0.674/-	-	-/-/0.270	0.384/-
CVPR2020	**Siam R-CNN**	0.701/0.891	0.649/0.834	0.609/0.220/0.408	**0.648/0.722**
	SiamBAN	0.696/0.910	0.631/0.833	0.597/0.178/0.452	0.514/0.598
	SiamAttn	**0.712/0.926**	**0.650/0.845**	**0.63/0.16/0.470**	0.560/0.648
	SiamCAR	-	0.614/0.760	-	0.507/0.510

4. KẾT LUẬN

Bài báo đã trình bày các chỉ số dùng để đánh giá các trình theo dõi đối tượng đơn, cùng với đó là các bộ dữ liệu với các thông số và thách thức tương ứng cũng được khảo sát. Mạng nơron Siamese gồm hai hoặc nhiều nhánh để đo lường mức độ tương tự giữa mục tiêu mẫu và vùng tìm kiếm, được sử dụng rộng rãi trong các trình theo dõi trong thời gian gần đây. Kết quả chính của bài báo là trình bày sâu tám mô hình theo dõi sử dụng kiến trúc Siamese được đề xuất từ năm 2018 đến năm 2020. So sánh đánh giá kết quả thực nghiệm của các trình theo dõi này trên các bộ dữ liệu phổ biến OTB2015, UAV123, VOT2018 và LaSOT đã được đề cập. Kết quả cho thấy trình theo dõi SiamAttn tính toán tương quan giữa các đặc trưng chú ý có kết quả cao nhất trên các bộ dữ liệu OTB2015, UAV123, VOT2018 và SiamRCNN có kết quả cao nhất trên bộ dữ liệu LaSOT. Các mô hình đề xuất sau đã khắc phục được một số hạn chế của mô hình trước và có kết quả cao hơn trên một số bộ dữ liệu cụ thể.

TÀI LIỆU THAM KHẢO

[1]. C. Robin and S. Lacroix, "*Multi-robot target detection and tracking: Taxonomy and survey*," Autonomous Robots, vol. 40, no. 4, pp. 729-760, 2016.

[2]. K. Lee, J. Hwang, G. Okopal, and J. Pitton, "*Ground-moving-platform-based human tracking using visual SLAM and constrained multiple kernels*," IEEE Trans. Intell. Transp. Sys., vol. 17, no. 12, pp. 3602-3612, 2016.

[3]. M. Gao, L. Jin, Y. Jiang, and B. Guo, "*Manifold Siamese network: Anovel visual tracking ConvNet for autonomous vehicles*," IEEE Trans.Intell. Transp. Sys., pp. 1-12, 2019.

[4]. J. Hao, Y. Zhou, G. Zhang, Q. Lv, and Q. Wu, "*A review of target tracking algorithm based on UAV*," in Proc. IEEE CBS, 2019, pp. 328-333.

[5]. M. Fiaz, A. Mahmood, and S. K. Jung, "*Tracking noisy targets: A review of recent object tracking approaches*," 2018. [Online]. Available: http://arxiv.org/abs/1802.03098

[6]. M. Fiaz, A. Mahmood, S. Javed, and S. K. Jung, "*Handcrafted and deep trackers: Recent visual object tracking approaches and trends*," ACM Computing Surveys, 2019.

[7]. P. Li, D. Wang, L. Wang, and H. Lu, "*Deep visual tracking: Review and experimental comparison*," Pattern Recognit., vol. 76, pp. 323-338,2018.

[8]. Abbass, Mohammed Y., Ki-Chul Kwon, Nam Kim, Safey A. Abdelwahab, Fathi E. Abd El-Samie, and Ashraf AM Khalaf. "*A survey on online learning for visual tracking*," The Visual Computer (2020): 1-22.

[9]. R. Pflugfelder, "*An in-depth analysis of visual tracking with Siamese neural networks*," 2017. [Online]. Available: http://arxiv.org/abs/1707.00569.

[10]. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll´ar, P., Zitnick, C.L. "*Microsoft coco: Common objects in context,*". In ECCV 2014, pp. 740-755.

[11]. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "*ImageNet large scale visual recognition challenge*," IJCV, vol. 115, no. 3, pp. 211-252, 2015.

[12]. E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "*YouTube-BoundingBoxes: A large high-precision human-annotated data set forobject detection in video*," in Proc. IEEE CVPR, 2017, pp. 7464-7473.

[13]. Y. Wu, J. Lim, and M. Yang, "*Object tracking benchmark*," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1834-848, 2015.

[14]. M. Mueller, N. Smith, and B. Ghanem, "*A benchmark and simulator for UAV tracking*," in Proc. ECCV, 2016, pp. 445-461.

[15]. M. Kristan, A. Leonardis, J. Matas, M. Felsberg, R. Pflugfelder, and et al., "*The sixth visual object tracking VOT2018 challenge results*," in Proc. ECCVW, 2018, pp. 3-53.

[16]. H. Fan, L. Lin, F. Yang, P. Chu, G. Deng, S. Yu, H. Bai, Y. Xu, C. Liao, and H. Ling, "*LaSOT: A high-quality benchmark for large-scale single object tracking*,". IEEE CVPR, 2019.

[17]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "*ImageNet classification with deep convolutional neural networks*," in Proc. NIPS, vol. 2, 2012, pp. 1097-1105.

[18]. K. He, X. Zhang, S. Ren, and J. Sun, *"Deep residual learning for image recognition,"* in Proc. IEEE CVPR, 2016, pp. 770-778.

[19]. B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "*High performance visual tracking with Siamese region proposal network*," in Proc. IEEE CVPR, 2018, pp. 8971–8980.

[20]. Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "*Distractor-aware Siamese networks for visual object tracking*," in Proc. ECCV, vol. 11213 LNCS, 2018, pp. 103-119.

[21]. B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "*SiamRPN++: Evolution of Siamese visual tracking with very deep networks*,". IEEE CVPR, 2019.

[22]. Z. Zhang and H. Peng, "*Deeper and wider Siamese networks for real-time visual tracking*,". IEEE CVPR, 2019.

[23]. Y. Yu, Y. Xiong, W. Huang, and M. R. Scott, "*Deformable siamese attention networks for visual object tracking*," in Proc. IEEE CVPR, 2020.

[24]. Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "*Siamese box adaptive network for visual tracking,*" in Proc. IEEE CVPR, 2020.

[25]. D. Guo, J. Wang, Y. Cui, Z. Wang, and S. Chen, "*SiamCAR: Siamese fully convolutional classification and regression for visual tracking*," in Proc. IEEE CVPR, 2020.

[26]. P. Voigtlaender, J. Luiten, P. H. Torr, and B. Leibe, "*SiamRCNN: Visual tracking by re-detection,*" in Proc. IEEE CVPR, 2020.

ABSTRACT

STUDY OF VISUAL OBJECT TRACKING METHODS
BASED ON SIAMESE NEURAL NETWORK ARCHITECTURE

Visual object tracking has become one of the most active research topics in computer vision, which has been growing in commercial development as well as academic research. Deep learning technologies have significantly advanced the task of visual object tracking, by providing the strong capacity of learning powerful deep features. Siamese-based trackers have achieved excellent performance on visual object tracking. This survey presents a deep analysis of the learning and inference capabilities of eight trackers based on Siamese neural networks in recent years. Quantitative results from eight papers are compared on popular visual tracking benchmarks OTB2015, UAV123, VOT2018, LaSOT.

Keywords: Computer Vision, Object Tracking, Deep Learning, Siamese Neural Networks.

Nhận bài ngày 06 tháng 4 năm 2021
Hoàn thiện ngày 20 tháng 8 năm 2021
Chấp nhận đăng ngày 05 tháng 9 năm 2021

Địa chỉ: [1]Viện Công nghệ thông tin, Viện Khoa học và Công nghệ quân sự;
[2]Trường Đại học Công nghệ, Đại học Quốc gia Hà Nội.
*Email: hoangdinhthang@gmail.com.