Xây dựng bộ dữ liệu UIT-DET cho hệ thống giám sát giao thông trong thành phố thông minh

Võ Hoàng Thông[1,2], Nguyễn Ngân Linh[1,2], Nguyễn Thiên Long[1,2], Võ Đình Tứ[1,2], Đỗ Trọng Hợp[1,2]

[1] Trường Đại học Công nghệ Thông tin
[2] Đại học Quốc gia Thành phố Hồ Chí Minh
{18521462, 18520989, 18521046, 18521589}@gm.uit.edu.vn, hopdt@uit.edu.vn

Tóm tắt nội dung Hệ thống giám sát quản lý giao thông là một trong những nền tảng xây dựng nên thành phố thông minh. Trong đó, để hệ thống này được vận hành đòi hỏi một lượng lớn dữ liệu thực tế các phương tiện giao thông ở các tình huống và điều kiện thời tiết khác nhau, và phân tích các hành vi chuyển động từ các phương tiện giao thông bằng các phương pháp phát hiện đối tượng. Trong công trình này, chúng tôi đề xuất một bộ dữ liệu mới UIT-DET cho các hệ thống quản lý giao thông trong thành phố thông minh và đánh giá hiệu suất của các hệ thống này dựa trên các chỉ số đánh giá mô hình phát hiện đối tượng. Bộ dữ liệu UIT-DET bao gồm 16 đoạn video quay được từ các bối cảnh giao thông trong thế giới thực (với 8000 khung hình và 41741 bounding boxes được gán nhãn, bao gồm các điều kiện như mức độ chiếu sáng, các loại xe, tỷ lệ kích thước và các bouding boxes). Chúng tôi thực nghiệm mô hình phát hiện đối tượng state-of-the-art EfficientDet trên bộ dữ liệu UIT-DET và phân tích những yếu tố ảnh hưởng phức tạp đến hiệu suất của hệ thống quản lý giao thông nói chung và mô hình nói riêng.
Từ khóa: Xây dựng bộ dữ liệu các phương tiện giao thông, phát hiện các phương tiện giao thông, hệ thống giám sát giao thông trong thành phố thông minh.

Abstract. A traffic management monitoring system is one of the foundations for building a smart city. For this system to operate requires a large amount of real data of vehicles in different weather situations and analysis of movement behaviors from vehicles by object detection methods. In this work, we propose a new dataset UIT-DET for traffic management systems in smart cities and evaluate the performance of these systems is based on the object detection model evaluation metrics. The UIT-DET dataset consists of 16 videos captured from real-world traffic scenes (with 8000 frames and 41741 labeled bounding boxes, including conditions such as illumination levels, vehicle types, aspect ratio, and bounding boxes). We evaluate the state-of-the-art EfficientDet object detection model on the UIT-DET dataset and analyze the complex influences on the performance of the traffic management system in general and the model in particular.

Keywords: A traffic vehicles dataset, vehicle detection, and monitor traffic system in a smart city.

1 Giới thiệu

Với sự phát triển nhanh chóng của khoa học công nghệ từ những thập niên thứ 10 của thế kỉ 21, xuất phát từ khái niệm cách mạng công nghiệp 4.0 trong một báo cáo chiến lược của chính phủ Đức đề cập tới những trụ cột chính của công nghệ, bao gồm: dữ liệu lớn (Big Data), người máy tự động (Autonomous Robots), công nghệ kết nối vạn vật (The Industrial Internet of Things - IoT) [1] mà nổi trội trong số đó được đông đảo các nhà khoa học quan tâm là bài toán về hệ thống quản lý giao thông trong thành phố thông minh (Traffic Management System). Việc phát triển thành công hệ thống này sẽ là một trong những tiền đề để xây dựng thành phố thông minh (Smart City). Bên cạnh đó, khả năng tận dụng camera giao thông với mạng lưới cảm biến phủ khắp đô thị trong việc tối ưu hóa luồng và quản lý phân chia các tuyến đường giao thông tránh tình trạng kẹt xe, kết nối và luân chuyển giao thông liên tục giữa các khu vực trong thành phố. Điều mà chúng ta đang thiếu là khả năng phát hiện và theo dõi các phương tiện giao thông qua các khu vực rộng lớn có nhiều camera ở những giao lộ khác nhau trong mọi điều kiện thời tiết và một lượng lớn dữ liệu các phương tiện giao thông được gán nhãn huấn luyện cho các mô hình phát hiện đối tượng. Để đạt được mục tiêu này, ta phải giải quyết các vấn đề nghiên cứu riêng biệt nhưng có mối liên hệ chặt chẽ với nhau: Xây dựng bộ dữ liệu kích thước lớn các phương tiện giao thông được gán nhãn cho mô hình phát hiện đối tượng và triển khai thuật toán phát hiện đa đối tượng.

Trong bài báo này, chúng tôi đề xuất bộ dữ liệu kích thước lớn (UIT-DET) cho các mô hình phát hiện các phương tiện giao thông cho hệ thống giao thông trong thành phố thông minh. Những đóng góp chính được đề cập trong công trình này được tóm tắt như sau:

- Bộ dữ liệu UIT-DET bao gồm 16 videos và hơn 8000 khung hình (frames) từ bối cảnh giao thông tại thành phố lớn. Các videos được gán nhãn thủ công với tổng số 41741 bouding boxes các phương tiện giao thông và các thuộc tính liên quan ví dụ như độ sáng của các bối cảnh, các loại xe và sự tắc nghẽn giao thông (xem Bảng 1).
- Chúng tôi xác định và đánh giá phương pháp thử nghiệm trên tập dữ liệu UIT-DET gồm mô hình state of the art của bài toán phát hiện đối tượng trong thị giác máy tính.

2 Các công trình nghiên cứu liên quan

Nhiều bộ dữ liệu điểm chuẩn được xây dựng cho bài toán phát hiện đa đối tượng như Caltech (Dollár và cộng sự, 2012) [2], KITTI-D (Geiger và cộng sự, 2012)

[3], PASCAL VOC (Everingham và cộng sự, 2015) [4], ImageNet (Russakovsky và cộng sự, 2015)[5], và KAIST (Hwang và cộng sự, 2015) [6]. Các bộ dữ liệu này chủ yếu được phát triển để huấn luyện cho các mô hình phát hiện đối tượng trong các hình ảnh đơn lẻ, ngoài ra còn có thể được sử dụng để huấn luyện các mô hình phát hiện đối tượng tự động cho các hệ thống phát hiện đối tượng. Bên cạnh đó, The Oxford Robotic Car (Maddern và cộng sự, 2017) [7] là một bộ dữ liệu xe tự hành với khoảng 20 triệu hình ảnh LIDAR, GPS và INS kèm nhãn dữ liệu trong mọi điều kiện thời tiết. Ngoài ra còn có bộ dữ liệu The Baidu Apolloscapes (Huang và cộng sự, 2018) [8] cung cấp bản đồ đám mây tọa độ điểm 3D, mỗi điểm ảnh, mỗi khung hình được đánh nhãn ngữ nghĩa, nhãn dấu làn đường và chú thích phân đoạn. Trong những năm gần đây, một bộ dữ liệu được phát triển giúp các nhà khoa học thuận lợi hơn trong việc tiếp cận lĩnh vực xe tự hành, là hệ thống cơ sở dữ liệu the Berkeley DeepDrive BDD100k (Yu và cộng sự, 2018) [9] với hơn 100,000 videos kèm nhãn gán của ảnh. Và cuối cùng là bộ dữ liệu The UA-DETRAC (Longyin Wen và cộng sự, 2020) [10], bộ dữ liệu cho bài toán phát hiện các phương tiện giao thông và phục vụ cho việc đánh giá các hệ thống theo dõi đa mục tiêu, với 100 videos, hơn 10 giờ quay, gần 140000 frames, 8250 loại xe cộ được gán nhãn và 1.21 triệu bounding boxes các phương tiện giao thông được đánh nhãn.

So với các tập dữ liệu xây dựng cho bài toán phát hiện và theo dõi đa đối tượng hiện có, tập dữ liệu UIT-DET được thiết kế cho các tình huống giám sát phương tiện giao thông với nhiều khung hình video, các bounding boxes được gán nhãn và các thuộc tính đa dạng. Sự khác biệt giữa bộ dữ liệu UIT-DET so với các bộ dữ liệu khác hiện có và đề xuất ở nhiều khía cạnh khác nhau được tóm tắt trong bảng 1 .

Bảng 1: Tóm tắt các bộ dữ liệu cho các bài toán phát hiện và truy vết đối tượng hiện có. Sáu cột đầu tiên là số lượng dữ liệu huấn luyện/kiểm thử (1k=1000) cho biết số lượng hình ảnh có chứa ít nhất một đối tượng, số lượng các vật thể được gán nhãn cho bài toán theo vết và số lượng các bounding boxes. Các cột còn lại là thuộc tính bổ sung của dữ liệu, "D": Bài toán phát hiện đối tượng, "T": Bài toán theo vết đối tượng, "P": Đối tượng trong bộ dữ liệu là người đi bộ và "C": Đối tượng trong bộ dữ liệu là phương tiện giao thông .

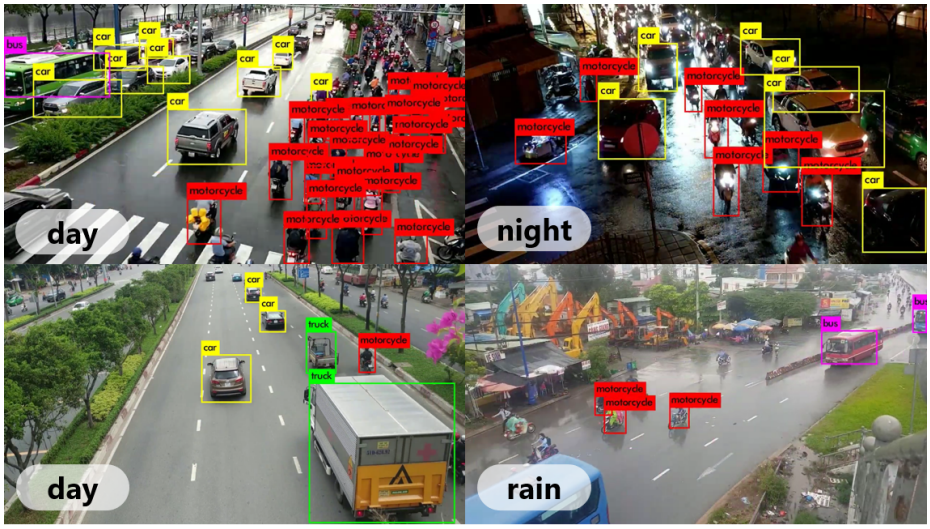
Bộ dữ liệu	Tập huấn luyện			Tập kiểm thử			Thuộc tính					
	Khung hình	Tracks	Bounding boxes	Khung hình	Tracks	Bounding boxes	Màu sắc	Video	Bài toán	Đối tượng	Đa dạng điều kiện chiếu sáng	Năm
INRIA [11]	1.2k	-	1.2k	741	-	566	✓		D	P		2005
ETH [12]	490	-	1.6k	1.8k	-	9.4k	✓	✓	D	P		2007
NICTA [13]	-	-	18.7k	-	-	6.9k	✓		D	P		2008
TUD-B [14]	1.09k	-	1.8k	508	-	1.5k	✓		D	P		2009
Caltech [2]	67k	-	192k	65k	-	155k	✓		D	P		2012
CUHK [15]	-	-	-	1.06k	-	-	✓	✓	D	P		2012
KITTI-D [3]	7.48k	-	40.6k	7.52k	-	39.7k	✓		D	P, C		2014
KAIST [6]	50.2k	-	41.5k	45.1k	-	44.7k	✓	✓	D	P	✓	2015
UIT-DET	7.2k	-	37.4k	800	-	4.1k	✓	✓	D	C	✓	2020
TUD [16]	610	-	610	451	31	2.6k	✓	✓	D,T	P		2008
PETS2009 [17]	-	-	-	1.5k	106	18.5k	✓	✓	D,T	P	✓	2009
UA-DETRAC [10]	84k	5.9k	578k	56k	2.3k	632k	✓	✓	D,T	C	✓	2015

3 Bộ dữ liệu UIT-DET

Bộ dữ liệu UIT-DET bao gồm 16 videos và 1,5 giờ chuỗi hình ảnh từ 11 địa điểm khác nhau, với các mô hình và điều kiện giao thông khác nhau bao gồm đường cao tốc đô thị, các điểm giao cắt và các nút giao thông. Đáng chú ý, để đảm bảo tính đa dạng, chúng tôi thu thập dữ liệu tại các vị trí, điều kiện ánh sáng và góc chụp khác nhau. Các video được ghi ở tốc độ 30 khung hình / giây (fps) và 10 khung hình / giây (fps) với độ phân giải hình ảnh 1920 x 1080 và 1280 x 720.

3.1 Thu thập và gán nhãn dữ liệu

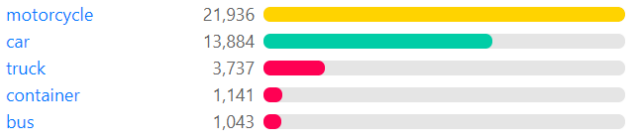
Gán nhãn video: Với 8000 khung hình trong bộ dữ liệu UIT-DET được đánh nhãn theo cách tổ chức dữ liệu của mô hình YOLO [18] với 41741 phương tiện giao thông. Dữ liệu thu thập được gán bởi 4 sinh viên chuyên ngành Khoa học Dữ liệu trong hơn hai tháng và được chúng tôi kiểm tra chéo để đảm bảo chất lượng của các chuỗi hình ảnh được gán. Tương tự như bộ dữ liệu PASCAL VOC (Everedham và cộng sự, 2015)[4], có một số vùng bị loại bỏ trong mỗi khung hình, bao gồm các phương tiện không thể được gán nhãn do độ phân giải thấp. Hình 1 cho thấy các khung hình với các lớp được gán nhãn trong tập dữ liệu UIT-DET. Các nhãn gán trong hình được phân tích ở các điều kiện khác nhau, ví dụ điều kiện chiếu sáng ban ngày, lớp car (xe hơi) bị che khuất bởi khu vực khác hay mật độ đông đúc khiến các phương tiện che khuất lẫn nhau. Điều kiện chiếu sáng được biểu thị bằng văn bản ở mỗi góc trái của khung hình lần lượt là ban ngày, ban đêm và trời mưa.



Hình 1: Các khung hình với các lớp được gán nhãn trong tập dữ liệu UIT-DET.

Tập dữ liệu UIT-DET được chia thành 3 tập, tập huấn luyện (UIT-DET-train), tập hiệu chỉnh (UIT-DET-validation) và tập kiểm thử (UIT-DET-test) được chia theo tỉ lệ tương ứng là 0.7, 0.2 và 0.1 trên tập dữ liệu UIT-DET. Chúng tôi chọn trong tập dữ liệu huấn luyện được quay tại các địa điểm khác nhau từ tập kiểm thử và các video này có thuộc tính và điều kiện giao thông tương đồng nhau. Điều này làm giảm khả năng các mô hình bị quá khớp trong các tình huống đặc biệt. Ngoài ra, thuật toán đều được huấn luyện trên bộ dữ liệu UIT-DET-train, hiệu chỉnh trên tập UIT-DET-validation và được đánh giá trên tập UIT-DET-test.

Bộ dữ liệu UIT-DET chứa các videos có nhiều sự khác biệt về hình dáng, tỉ lệ, hình nền, độ chiếu sáng và mật độ giao thông. Bộ dữ liệu này phục vụ cho việc đánh giá khả năng phát hiện đối tượng, tương tự như bộ dữ liệu KITTI detection (Geiger và cộng sự, 2012) [3] và bộ dữ liệu WIDER FACE (Yanget và cộng sự, 2016) [19].



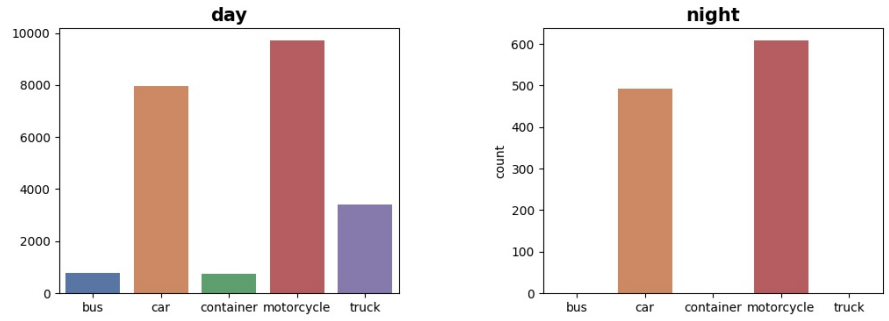
Hình 2: Phân phối số mẫu trong các lớp motorcycle, car, bus, truck, container

Để phân tích hiệu suất của các thuật toán phát hiện đối tượng, chúng tôi dựa trên độ phức tạp của một chuỗi các hình ảnh với các thuộc tính sau:

- Loại phương tiện giao thông: Chúng tôi gán nhãn năm loại phương tiện giao thông là motorcycle (xe máy, xe đạp điện, xe máy điện), car (xe hơi, xe taxi, xe bán tải), bus (xe buýt, xe khách), truck (xe tải, xe bồn, xe ben), container. Các loại phương tiện giao thông này được chúng tôi định nghĩa cụ thể như trên. Biểu đồ phân phối của các phương tiện giao thông trong hình 2.
- Độ chiếu sáng: Chúng tôi xem xét bốn loại điều kiện chiếu sáng, là ban đêm, trời nắng, trời mưa và điều kiện lóa sáng quá mức của các phương tiện giao thông có cường độ ánh sáng đèn xe ban đêm mạnh(glare). Biểu đồ phân phối các lớp thể hiện mức độ chiếu sáng của tập dữ liệu trong hình 3.
- Tỉ lệ: Chúng tôi xác định tỉ lệ các bouding box được gán nhãn của các phương tiện giao thông bằng căn bậc hai của vùng gán nhãn các điểm ảnh. Kích thước các nhãn xe được gán với tỉ lệ như sau: tỉ lệ nhỏ (0-32 pixels), tỉ lệ trung bình (32-96 pixels) và tỉ lệ lớn (hơn 96 pixels).

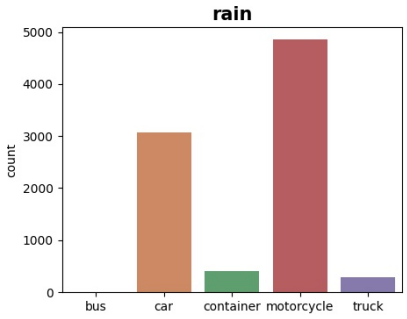
3.2 Thống kê tập dữ liệu

Trong phần này, chúng tôi phân tích các thuộc tính và khía cạnh của bộ dữ liệu UIT-DET, bộ dữ liệu được xây dựng chủ yếu cho lớp bài toán phát hiện đối

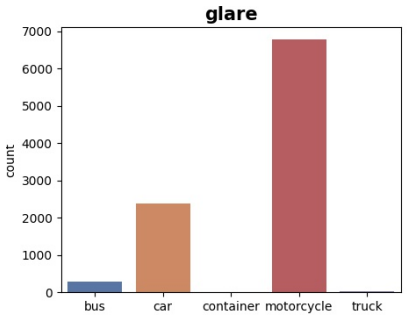


(a) Ban ngày

(b) Ban đêm



(c) Trời mưa



(d) Cường độ ánh sáng mạnh

Hình 3: Phân phối số lượng phương tiện ở các điều kiện ánh sáng và thời tiết

tượng nói chung hay các hệ thống giao thông cho thành phố thông minh để phát hiện các phương tiện giao thông nói riêng.

Số lượng mẫu mỗi lớp trong tất cả 5 lớp được thể hiện trong hình 2. Số lượng mẫu mỗi lớp trong từng điều kiện thời tiết khác nhau được thể hiện trong hình 3. Số lượng lớp motorcycle chiếm đa số trong tất cả các lớp và cũng là lớp có số lượng nhãn gán mà không có bộ dữ liệu nào có, hay được xem như điểm đặc biệt của bộ dữ liệu UIT-DET so với các bộ dữ liệu được phát triển cho bài toán phát hiện đối tượng. Trong bộ dữ liệu, một số đối tượng có kích thước nhỏ nên khó xác định và đòi hỏi nhiều lý luận về ngữ cảnh và điều kiện chiếu sáng để có thể nhận ra.

4 Phương pháp tiếp cận

Có 2 cách tiếp cận chính phổ biến và hiện đại nhất đối với các hệ thống hay thiết bị dùng để phát hiện đối tượng là two-stage và single-stage detectors.

4.1 Two-Stage Detectors

Two-Stage Detectors là một phương pháp tiếp cận để giải quyết bài toán phát hiện đối tượng, với ý tưởng là sử dụng mạng đề xuất khu vực (a Region Proposal NetWork) để đưa ra các vị trí giả định mà đối tượng có thể tồn tại trong ảnh, sau đó các vùng đề xuất này được phân lớp và đưa ra các dự báo hồi quy các bounding box của những vật thể. Một ví dụ điển hình dựa trên ý tưởng này là Faster R-CNN. Faster R-CNN (Ren và các cộng sự, 2017) [20] đã đưa ra giải pháp sử dụng cơ chế với mạng lưới đề xuất khu vực (a region proposal network - RPN) nhằm cải thiện các phương pháp được đề xuất trước đó (Girshick, 2015; He và cộng sự, 2015).

Tương tự như Fast R-CNN [21], đầu tiên hình ảnh được đưa vào một mạng tích chập (CNN), đầu ra trả về là một bản đồ trích xuất đặc trưng (feature map) của mạng tích chập, ghi lại kết quả của việc áp dụng các bộ lọc cho hình ảnh đầu vào, tức là với mỗi lớp bản đồ trích xuất đặc trưng là kết quả đầu ra của lớp đó. Thay vì áp dụng tìm kiếm có chọn lọc (selective search) trên bản đồ đặc trưng để phát hiện các khu vực đề xuất, một mạng lưới riêng biệt được xây dựng để dự đoán các khu vực được đề xuất. Các vùng đề xuất được dự đoán sau đó được biến đổi lại bằng cách đưa qua lớp ROI pooling và cuối cùng sẽ thực hiện phân loại hình ảnh dựa trên các vùng được đề xuất đồng thời dự đoán các giá trị offset cho các bounding box.

4.2 One-Stage Detectors

One-Stage Detectors là một phương pháp tiếp cận cho bài toán phát hiện đối tượng, với ý tưởng đưa bài toán về dạng bài toán hồi quy bằng cách từ hình ảnh đầu vào, mô hình đưa ra đồng thời các dự đoán xác suất cho biết đối tượng thuộc lớp nào và dự báo tọa độ các vật thể có trong ảnh đầu vào. Điển hình trong cách tiếp cận này là họ mô hình phát hiện đối tượng YOLO.

Mô hình You-Only-Look-Once(YOLO) được đề xuất bởi Redmon và cộng sự (2016) [18], là một thuật toán phát hiện đối tượng theo thời gian thực, với ý tưởng đưa về dạng mô hình hồi quy để dự đoán vật thể. YOLO sử dụng một kiến trúc mạng nơ-ron duy nhất để dự đoán các bounding box và xác suất phân lớp từ hình ảnh.

Trong những năm gần đây, một mô hình đạt hiệu suất cao trên các bộ dữ liệu điểm chuẩn là EfficientDet được đề xuất dựa trên mạng BiFPN do Mingxing Tan, Ruoming Pang, và Quoc V. Le đến từ Google Research Brain Team phát triển [22]. Mô hình EfficientDet sử dụng mạng xương sống (backbone network) EfficientNet, mạng trích xuất đặc trưng BiFPN, và một mạng để dự đoán lớp và bounding box. Tất cả lớp từ BiFPN và các lớp trong kiến trúc mô hình được dùng để dự đoán lớp và bounding box đều được học bằng cách lặp đi lặp lại nhiều lần dựa trên những ràng buộc khác nhau về mặt tài nguyên.

5 Thực nghiệm và phân tích

Trong phần này, chúng tôi sẽ giới thiệu cách thức đánh giá cho bài toán phát hiện đối tượng và thực nghiệm trên mô hình state-of-the-art EfficientDet.

5.1 Thông số đánh giá

Chúng tôi dựa trên chỉ số Average precision (AP), chỉ số trên được tính theo đường cong precision vs. recall(PR) để đánh giá cho mỗi thuật toán phát hiện đối tượng. Đường cong Precision-Recall được tạo ra bằng cách vẽ từng điểm có toạ độ trên trục toạ độ và nối chúng với nhau.

5.2 Kết quả đánh giá thực nghiệm

Hiệu suất: Trong Bảng 2, chúng tôi phân tích và so sánh hiệu suất của phương pháp phát hiện đối tượng trên bộ dữ liệu UIT-DET dựa trên các chỉ số đánh giá mean average precision (mAP) và average precision (AP). Thông thường, với bài toán phát hiện đối tượng, mAP là chỉ số chính được dùng để đánh giá.

Kết quả của mô hình state-of-the-art trên bộ dữ liệu UIT-DET, được hiển thị trong bảng 2 với độ đo đánh giá mAP. Cụ thể, phương pháp EfficientDet hoạt động khá tốt với 0.783 và 0.785 mAP với ngưỡng IoU > 0.5.

Tỉ lệ: Bảng 3 thể hiện kết quả đánh giá tỉ lệ kích thước phương tiện giao thông trong tập kiểm thử UIT-DET. Đối với phương tiện có kích thước nhỏ, là những phương tiện có điểm ảnh bé hơn kích thước 32x32, phương pháp EfficientDet nhận diện cực kỳ kém với AP là 0.00 %. Đối với phương tiện có kích thước trung bình, những phương tiện có kích thước điểm ảnh từ 32 x 32 đến 96 x 96, mô hình EfficientDet cho kết quả là 43.4%. Với phương tiện có kích thước lớn, kích thước điểm ảnh của phương tiện trên 96 x 96, thì mô hình EfficientDet lại cho kết quả tốt là 85.1%.

Bảng 2: *Phương pháp phát hiện đối tượng dựa trên các chỉ số đánh giá mAP*

Mô hình	mAP@0.5	mAP@0.75	mAP@0.5:0.95
EfficientDet	0.785	0.762	0.681

Bảng 3: *Chỉ số AP dựa trên kích thước các phương tiện giao thông nhỏ, trung bình và lớn của mô hình EfficientDet*

Mô hình	AP small	AP medium	AP large
EfficientDet	0.00%	43.4%	85.1%

6 Kết luận

Trong bài báo này, chúng tôi trình bày bộ dữ liệu kích thước lớn dùng để huấn luyện các mô hình phát hiện đối tượng cho các hệ thống giám sát giao thông trong thành phố thông minh gồm 16 video với 8000 khung hình đã được gán nhãn và 41741 bounding boxes. Bên cạnh đó, chúng tôi thực hiện các thử nghiệm trên mô hình phát hiện đối tượng state-of-the-art EfficientDet để đánh giá hiệu suất dựa trên các chỉ số. Qua đó cho thấy sự ảnh hưởng của các yếu tố lên độ chính xác của mô hình phát hiện đối tượng, đặc biệt là trong các hệ thống giám sát giao thông.

Lời cảm ơn

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ Đề tài mã số C2021-26-04.

Tài liệu

[1] G. Erboz, "How to define industry 4.0: main pillars of industry 4.0", in (Nov. 2017).

[2] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art", IEEE Transactions on Pattern Analysis and Machine Intelligence **34**, 743–761 (2012).

[3] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite", in 2012 ieee conference on computer vision and pattern recognition (2012), pp. 3354–3361.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[5] Russakovsky, Olga, and et al, "Imagenet large scale visual recognition challenge", CoRR **abs/1409.0575** (2014).
[6] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: benchmark dataset and baseline", in 2015 ieee conference on computer vision and pattern recognition (cvpr) (2015), pp. 1037–1045.
[7] Maddern, Will, and et al, "1 Year, 1000km: The Oxford RobotCar Dataset", The International Journal of Robotics Research (IJRR) **36**, 3–15 (2017).
[8] Huang, Xinyu, and et al, "The apolloscape dataset for autonomous driving", CoRR **abs/1803.06184** (2018).
[9] Yu, Fisher, and et al, "BDD100K: A diverse driving video database with scalable annotation tooling", CoRR **abs/1805.04687** (2018).
[10] Wen, Longyin, and et al, "DETRAC: A new benchmark and protocol for multi-object tracking", CoRR **abs/1511.04136** (2015).
[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", in 2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05), Vol. 1 (2005), 886–893 vol. 1.
[12] A. Ess, B. Leibe, and L. Van Gool, "Depth and appearance for mobile scene analysis", in 2007 ieee 11th international conference on computer vision (2007), pp. 1–8.
[13] G. Overett, L. Petersson, N. Brewer, L. Andersson, and N. Pettersson, "A new pedestrian dataset for supervised learning", in 2008 ieee intelligent vehicles symposium (2008), pp. 373–378.
[14] C. Wojek, S. Walk, and B. Schiele, "Multi-cue onboard pedestrian detection", in 2009 ieee conference on computer vision and pattern recognition (2009), pp. 794–801.
[15] W. Ouyang and X. Wang, "A discriminative deep model for pedestrian detection with occlusion handling", in 2012 ieee conference on computer vision and pattern recognition (2012), pp. 3258–3265.
[16] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking", in 2008 ieee conference on computer vision and pattern recognition (2008), pp. 1–8.
[17] J. Ferryman and A. Shahrokni, "Pets2009: dataset and challenge", in 2009 twelfth ieee international workshop on performance evaluation of tracking and surveillance (2009), pp. 1–6.
[18] Redmon, Joseph, and et al, "You only look once: unified, real-time object detection", CoRR **abs/1506.02640** (2015).
[19] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: a face detection benchmark", in 2016 ieee conference on computer vision and pattern recognition (cvpr) (2016), pp. 5525–5533.
[20] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks", CoRR **abs/1506.01497** (2015).
[21] R. B. Girshick, "Fast R-CNN", CoRR **abs/1504.08083** (2015).
[22] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: scalable and efficient object detection", CoRR **abs/1911.09070** (2019).