Tạp chí Khoa học Trường Đại học Cần Thơ
Phần A: Khoa học Tự nhiên, Công nghệ và Môi trường
website: ctujsvn.ctu.edu.vn

DOI:10.22144/ctu.jvn.2022.159

THỰC NGHIỆM ĐÁNH GIÁ DOUBLE-HEAD CHO BÀI TOÁN PHÁT HIỆN PHƯƠNG TIỆN GIAO THÔNG TỪ KHÔNG ẢNH

Nguyễn Thanh Thanh Trúc*, Trần Thị Mỹ Quyên, Bùi Cao Doanh, Võ Duy Nguyên và Nguyễn Tấn Trần Minh Khang

Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh

**Người chịu trách nhiệm về bài viết: Nguyễn Thanh Thanh Trúc (email: 20520829@gm.uit.edu.vn)*

Thông tin chung:
Ngày nhận bài: 03/03/2022
Ngày nhận bài sửa: 21/04/2022
Ngày duyệt đăng: 25/04/2022

Title:
An emperical study of Double-Head for vehicle detection in aerial images

Từ khóa:
Học sâu, máy bay không người lái, phát hiện phương tiện giao thông, thành phố thông minh

Keywords:
Deep learning, smart city, unmaned aerial vehicle (UAVs), vehicle detection

ABSTRACT

Vehicle detection in aerial images problem poses multiple challenges and has been of great interest to many in the research community. Objects in aerial images are a lot smaller in size compared to those in images taken from the ground, which is one of the biggest challenges in this problem. With small objects, the differences between regional proposals gravely affect the detection result. In this research, the Double-Head method is evaluated on the AERIAU dataset, an aerial image dataset that utilizes data augmentation techniques. The Double-Head achieved an mAP score of 37.09% on the AERIAU dataset. Compared with the previous method that achieved the highest result on the AERIAU dataset, which wasYOLOv3, Double-Head was surpassed by 2.01%. The Double-Head model achieved remarkably high results in the 'car', 'bus', and 'truck' vehicle classes, from which proposals are made to detect smaller vehicles. This is a premise of future research and a basis for developing smart traffic surveillance systems.

TÓM TẮT

Phát hiện phương tiện giao thông từ không ảnh đặt ra nhiều thách thức và nhận được sự quan tâm từ cộng đồng nghiên cứu. Đối tượng trong không ảnh nhỏ hơn rất nhiều so với ảnh chụp từ camera mặt đất, đây là thách thức rất lớn. Với các đối tượng nhỏ, sự sai khác của các vùng đề xuất sẽ làm ảnh hưởng lớn đến kết quả phát hiện đối tượng. Trong nghiên cứu này, phương pháp Double-Head được đánh giá dựa trên bộ dữ liệu AERIAU – một bộ dữ liệu không ảnh có áp dụng các kỹ thuật tăng cường dữ liệu. Double-Head đạt kết quả 37,09% mAP trên bộ dữ liệu AERIAU. So sánh với mô hình đạt kết quả cao nhất được công bố trước đó trên bộ dữ liệu AERIAU là YOLOv3, Double-Head cao hơn 2,01%. Double-Head đạt kết quả cao trên lớp đối tượng xe ô tô, xe buýt, xe tải, từ đó đưa ra đề xuất phát hiện xe loại nhỏ. Đây là tiền đề cho các nghiên cứu tiếp theo, cơ sở để phát triển các hệ thống giám sát giao thông thông minh.

1. GIỚI THIỆU

Việt Nam là một trong những nước có tình hình giao thông phức tạp nhất trên toàn thế giới với số lượng phương tiện giao thông lưu thông trên đường hằng ngày đạt con số rất lớn. Theo số liệu của Cục Cảnh sát giao thông, tháng 7/2020 (Tổng cục Đường bộ Việt Nam, 2021), thành phố Hồ Chí Minh có 8.94 triệu phương tiện cá nhân, tăng gần 7% so với cùng kỳ năm 2018. Trong đó, có hơn 825.000 ô tô (tăng

gần 16%) và 8,12 triệu xe máy (tăng hơn 6%). Chỉ trong khoảng 10 năm (từ năm 2010 đến nay), phương tiện giao thông đã tăng thêm hơn 4 triệu. Theo thống kê, bình quân mỗi tháng có 30.000 phương tiện giao thông đăng ký mới, đồng nghĩa với mỗi ngày có 1.000 phương tiện đăng ký mới. Điều đó tạo áp lực cho các hệ thống giám sát cũng như kiểm soát giao thông. Phát hiện phương tiện giao thông là một bài toán không còn xa lạ thuộc nhóm các bài toán phát hiện đối tượng và có nhiều ứng dụng trên thực tế, là tiền đề giúp phát triển hệ thống giám sát giao thông thông minh. Bài toán này nhận được sự quan tâm của các nhà khoa học, các hãng sản xuất công nghiệp lớn nhằm phát triển các hệ thống tự động và điều tiết giao thông. Việc tìm được vị trí các đối tượng giao thông từ không ảnh giúp phát hiện các bất thường, quản lý hoạt động của các nút giao thông một cách hiệu quả, toàn diện hơn. Quản lý giao thông thông minh là chìa khóa cho một thành phố thông minh.

Hình ảnh chụp từ thiết bị bay không người lái (flycam, drone) hay còn được gọi là không ảnh được sử dụng rộng rãi trong những năm gần đây. Cụ thể là trong lĩnh vực thị giác máy tính, tuy giám sát giao thông là hoạt động thường xuyên, cần dựa vào các hệ thống camera giám sát được thiết lập cố định nhưng vì không ảnh có độ phân giải cao, tầm nhìn bao quát nên các phương tiện giao thông dễ dàng được phát hiện. Tuy nhiên, do tầm nhìn và quy mô của không ảnh là khá lớn nên việc phát hiện các phương tiện giao thông còn gặp cản trở bởi nhiều đối tượng gây nhiễu như tòa nhà, cầu, cây xanh, thùng rác, ... Điều này tạo nhiều khó khăn cho việc phát hiện đối tượng quan tâm một cách chính xác. Đặc biệt, đa số các đối tượng trong không ảnh chiếm tỷ lệ nhỏ, cụ thể là đối tượng xe mô tô, đây là thách thức lớn trong cộng đồng thị giác máy tính hiện nay.

Phát hiện phương tiện giao thông là chủ đề không quá xa lạ trong cộng đồng thị giác máy tính, các nghiên cứu về chủ đề này được công bố nhiều trong nước cũng như quốc tế (Ho et al., 2020; Liu et al., 2020). Trên cơ sở kế thừa những công trình nghiên cứu khoa học trước đó, nghiên cứu này hướng đến việc phân loại phương tiện giao thông trong không ảnh dựa trên bộ dữ liệu AERIAU (Chung et al., 2020), các lớp đối tượng được xem xét trong phạm vi nghiên cứu là xe ô tô (car), xe tải (truck), xe buýt (bus) và xe mô tô (motor). Các phương tiện có hình dạng và kích thước khác nhau, các vùng đề xuất tìm được sẽ có sự đa dạng về kích thước, phù hợp cho mục tiêu nghiên cứu. Ở Hình 1, các phương tiện được xác định như xe mô tô có kích thước nhỏ, to hơn là ô tô, xe buýt, xe tải. Khi thay đổi tầm nhìn, bài toán trở nên khó hơn, dễ nhận nhầm giữa các đối tượng và khó phát hiện đối tượng nhỏ.

Hình 1. Minh họa bài toán phát hiện phương tiện giao thông

Đầu vào (ảnh bên trái) là ảnh các phương tiện giao thông được chụp từ trên cao, đầu ra (ảnh bên phải) là hộp giới hạn chứa các đối tượng được gán nhãn theo các lớp car – xe ô tô, truck – xe tải, bus – xe buýt, motor – xe mô tô

Trong nghiên cứu này, bài toán phát hiện phương tiện giao thông từ không ảnh và nâng cao hiệu quả phát hiện đối tượng trong không ảnh với hai tác vụ là phân lớp ảnh và hồi quy hộp giới hạn bằng một cách tiếp cận khác được giải quyết. Cụ thể, đóng góp chính của nghiên cứu là:

– Nghiên cứu nhằm khảo sát và đánh giá hai mô hình: 1) mô hình Double-Head, một phương pháp hồi quy tọa độ đối tượng bằng một đầu tích chập (conv-head) và phân loại đối tượng bằng một đầu kết nối đầy đủ (fc-head); 2) phương pháp GRoIE, đây là một mô-đun trích xuất đặc trưng đối tượng mới và hiệu quả hơn RoI Pooling.

– Đề xuất một cách kết hợp Double-Head và GRoIE, gọi là GRoIE Double Head giúp tận dụng cả hai ý tưởng của hai phương pháp.

– Thực nghiệm và đánh giá phương pháp Double-Head trên bộ dữ liệu AERIAU.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Phát hiện hai giai đoạn (two-stage detector) bao gồm hai công việc chính: tạo các đề xuất khu vực

(Tìm kiếm có chọn lọc trong R-CNN và Fast-RCNN hoặc mạng đề xuất khu vực (RPN) trong Faster RCNN) và phân loại đối tượng cho mỗi khu vực được đề xuất. Chẳng hạn, trong Faster RCNN (Ren et al., 2015), ở giai đoạn đầu, mạng đề xuất khu vực (RPN) tạo ra các vùng đề xuất, sau đó các vùng đề xuất này được sử dụng để phân loại đối tượng. Cấu trúc của phát hiện hai giai đoạn linh hoạt, phù hợp hơn cho phân loại theo vùng quan tâm. Điểm nổi bật của nó là cho độ chính xác cao hơn phát hiện một giai đoạn tuy nhiên chi phí tính toán lớn và tốc độ khá chậm. R-CNN (Ren et al., 2015), R-FCN (Dai et al., 2016) là các mô hình tiêu biểu sử dụng kiến trúc hai giai đoạn. Mask R-CNN (He et al., 2017) là phiên bản mở rộng của Faster R-CNN, cấu trúc của Mask R-CNN cũng không quá khác biệt so với cấu trúc của Faster R-CNN. Với mục tiêu giải quyết bài toán phân đoạn hình ảnh (Segmentation), giai đoạn đầu tiên, Mask R-CNN vẫn sử dụng mạng đề xuất khu vực để phân loại vật thể và tạo hộp giới hạn nếu có. Với giai đoạn thứ hai, thay vì rút trích đặc trưng sử dụng "RoI Pooling" như Faster R-CNN thì Mask R-CNN lại sử dụng "RoI Align" để thực hiện việc này. ResNet (He et al., 2016), mạng ra đời với số lớp lớn nhưng vẫn giải quyết được vấn đề"tiêu biến gradient" hoặc "bùng nổ gradient" khi huấn luyện, đạt được tỷ lệ lỗi (error rate) top-5 là 3,57%, có số lớp tăng đáng kể so với các mạng trước đây. Mô hình ResNet-50 (He et al., 2016) bao gồm 5 giai đoạn, mỗi giai đoạn có một khối tích chập và khối phần dư. Mỗi khối tích chập có 3 lớp tích chập và mỗi khối phần dư cũng có 3 lớp tích chập. ResNet-50 có hơn 23 triệu tham số huấn luyện.

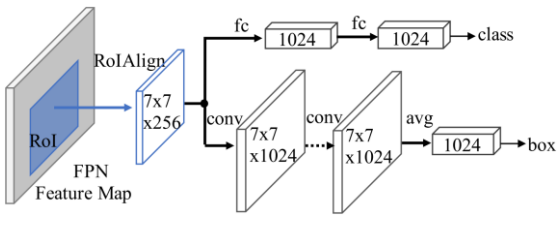
Phát hiện một giai đoạn (one-stage detector) trực tiếp bỏ qua giai đoạn đề xuất khu vực như phát hiện hai giai đoạn, chỉ bằng một giai đoạn trực tiếp nó có thể phân phối xác suất cho các lớp và định vị đối tượng trên ảnh. Phát hiện một giai đoạn hạn chế các thuật toán tiền xử lý tạo kiến trúc xương sống (backbone) gọn nhẹ và giảm thiểu vùng đề xuất được dự đoán. Nhờ vậy, tốc độ phát hiện đối tượng nhanh hơn nhưng độ chính xác thường thấp hơn so với phát hiện hai giai đoạn. Một số mô hình sử dụng kiến trúc một giai đoạn: YOLO (Redmon et al., 2016), SDD (Liu et al., 2016). YOLOv1 (Redmon et al., 2016) đánh dấu sự ra đời họ YOLO, lấy ý tưởng từ GoogLeNet (Szegedy et al., 2015).Nó hợp nhất các thành phần chuyên biệt tạo thành mạng nơ-ron duy nhất (24 lớp tích chập theo sau 2 lớp kết nối đầy đủ) với ảnh đầu vào có kích thước 224×224. YOLOv2 (Redmon et al., 2017) được tạo ra với mục tiêu khắc phục hạn chế của YOLOv1, YOLOv2 cải thiện độ chính xác và tăng tốc độ phát hiện. Thay vì có một kiến trúc nơ-ron phức tạp, phiên bản này biểu diễn thông tin đơn giản dễ học, trở thành "state-of-the-art" (67 FPS, 76,8% mAP trên VOC 2007) vượt trội so với Faster RCNN, ResNet và SSD. YOLOv3 (Redmon et al., 2018) có kiến trúc khá giống YOLOv2, tuy nhiên YOLOv3 đã có những thay đổi mới giúp cải thiện hiệu suất phát hiện các đối tượng nhỏ - đây cũng chính là nhược điểm của các phương pháp họ YOLO trước đó: sử dụng "hồi quy logistic" cho việc dự đoán độ tin cậy cho hộp giới hạn dựa trên ngưỡng cho trước (threshold = 0,5); dựa trên ý tưởng kim tử tháp tính năng (Feature Pyramid Networks), ứng với mỗi vị trí YOLOv3 đưa ra 3 dự đoán (hộp giới hạn, đối tượng, điểm số của lớp); Darknet-19 được thay thế bởi Darknet-53 thực thi tác vụ rút trích đặc trưng.

3. PHƯƠNG PHÁP NGHIÊN CỨU

3.1. Phương pháp Double-Head

Double-Head là phương pháp được đề xuất bởi Bae et al. (2020) để tận dụng lợi thế của cả hai đầu, bao gồm một đầu được kết nối đầy đủ (fc-head) để phân lớp ảnh và một đầu tích chập (conv-head) để hồi quy hộp giới hạn. Fc-head có độ nhạy nhất định về không gian (spatial sensitivity), fc-head có các tham số khác nhau cho các phần khác nhau của một đề xuất (proposal) và nhờ đó, mô hình có thể dễ dàng phân biệt các bộ phận hoặc những thành phần giữa các vật thể khác nhau nhưng fc-head không quá nổi trội trong việc xác định "miền offset" của toàn vật thể. Ngược lại, conv-head chia sẻ ma trận lọc tích chập cho tất cả các vị trí của bản đồ đặc trưng đầu vào và sử dụng "average pooling" để tổng hợp. Fc-head phù hợp hơn cho nhiệm vụ phân lớp vì sự chênh lệch điểm số phân loại ở các ngưỡng IoU của fc-head rõ rệt và hơn hẳn so với conv-head. Trong khi đó, conv-head hồi quy hộp giới hạn chính xác hơn.

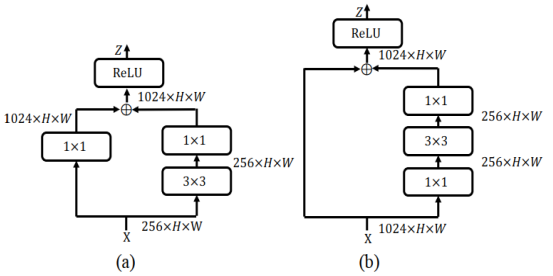
Phương pháp Double-Head chia phân loại ảnh và định vị hộp giới hạn thành fc-head và conv-head tương ứng, kiến trúc đầy đủ được thể hiện ở Hình 2. Chi tiết của kiến trúc xương sống và hai đầu mô hình Double-Head được mô tả như sau:



Hình 2. Minh họa kiến trúc Double-Head được Bae et al. (2020) đề xuất gồm một đầu kết nối đầy đủ và một đầu tích chập

– **Kiến trúc xương sống:** Sử dụng xương sống là FPN (Liu et al.,2016) để tạo ra các đề xuất khu vực (region proposals) và sử dụng "RoI Align" trích xuất đặc trưng đối tượng từ nhiều cấp độ. Mỗi đề xuất có một bản đồ đặc trưng với kích thước 256×7×7, được chuyển đổi bởi fc-head và conv-head thành hai vectơ đặc trưng (mỗi vectơ có kích thước 1024) để phân loại và hồi quy hộp giới hạn tương ứng.

– **Đầu kết nối đầy đủ (fc-head):** Có hai lớp kết nối đầy đủ. Kích thước đầu ra là 1024. Kích thước tham số là 13,25M.



Hình 3. Kiến trúc mạng của hai thành phần: (a) khối dư (residual block) tăng số kênh từ 256 lên 1024, (b) khối nút cổ chai (residual bottleneck block)

(Bae et al., 2021)

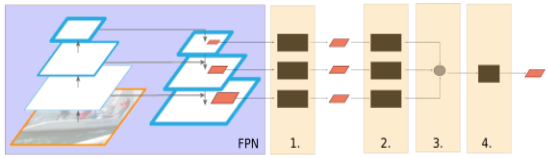
– **Đầu tích chập (conv-head):** Xếp chồng K khối dư (residual blocks). Khối đầu tiên tăng số lượng kênh từ 256 lên 1024 (Hình 3 (a)), và những khối khác là khối nút cổ chai (bottleneck blocks) (Hình 3 (b)). Cuối cùng, "average pooling" được sử dụng để tạo vectơ đặc trưng có kích thước 1024. Mỗi khối dư có 1,06 M tham số.

– **Loss Function:** Cả hai đầu (fc-head và conv-head) đều được huấn luyện với mạng đề xuất khu vực (RPN) từ đầu đến cuối. Tổn thất chung được tính như sau: $\mathcal{L} = \omega^{fc}\mathcal{L}^{fc} + \omega^{conv}\mathcal{L}^{conv} + \mathcal{L}^{rpn}$ trong đó ω^{fc} và ω^{conv} lần lượt là trọng số của fc-head và conv-head. $\mathcal{L}^{fc}$, $\mathcal{L}^{conv}$, $\mathcal{L}^{rpn}$ lần lượt là tổn thất đối với fc-head, conv-head và RPN.

3.2. Phương pháp GRoIE

GRoIE là phương pháp được cung cấp bởi Rossi et al. (2021). Phương pháp này trích xuất đặc trưng đối tượng từ bản đồ đặc trưng chung, nghĩa là nó sử dụng tất cả các lớp của FPN thay vì chỉ sử dụng một lớp (lớp tốt nhất) như trong các loại RoI truyền thống. Kiến trúc GRoIE cho phép chúng ta hưởng lợi từ thông tin có trong tất cả các lớp FPN, điều này giúp khắc phục được những hạn chế vốn có trong việc lựa chọn một lớp FPN duy nhất. Phương pháp này gồm 4 công việc chính: đầu tiên, thực hiện "max-pooling" trên vùng quan tâm không đồng nhất để được biểu diễn có kích thước cố định. Tiếp theo, các bản đồ đặc trưng được xử lý trước riêng biệt và sau đó được tổng hợp thành một bản đồ đặc trưng duy nhất. Cuối cùng, xử lý hậu kỳ được áp dụng để trích xuất thông tin. Phương pháp này phù hợp cho cả phát hiện đối tượng và phân đoạn cá thể.

Cấu trúc của GRoIE gồm có 4 phần tương ứng với 4 công việc chính đã đề cập ở trên: mô-đun "RoI pooler", mô-đun tiền xử lý, mô-đun tổng hợp và mô-đun hậu xử lý (Hình 4).



Hình 4. Cấu trúc GRoIE. (1) RoI Pooler (2) Preprocessing (3) Aggregation function (4) Post-processing

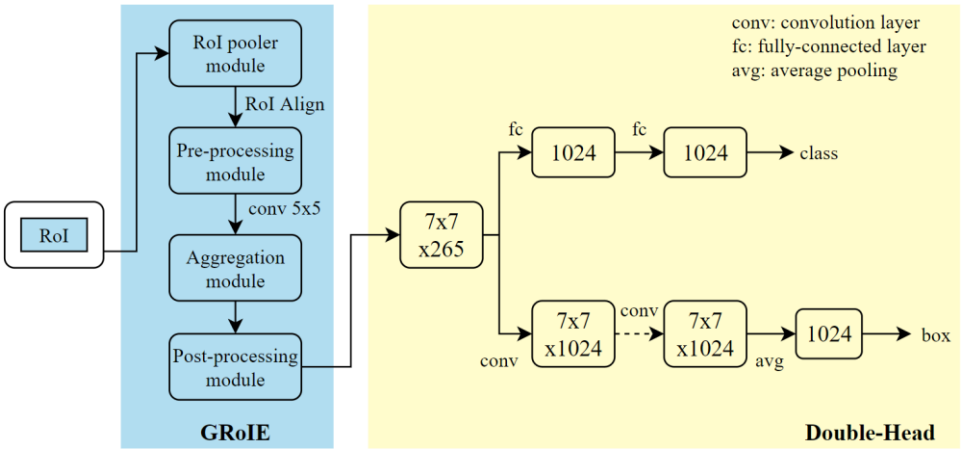
(Rossi et al., 2021)

– **Mô-đun "RoI pooler":** Từ đầu ra của RPN, trong mỗi quy mô sẽ có một RoI có kích thước cố định được sử dụng. Trong số các kỹ thuật "RoI pooling" hiện có, "RoI Align" (He et al., 2017) là thích hợp nhất vì nó làm giảm một vùng đặc trưng của đối tượng bằng cách chia RoI ban đầu trong các hộp bằng nhau và áp dụng nội suy song tuyến bên trong mỗi hộp. Điều này giúp tránh việc bị mất mát các phần tử ảnh.

– **Mô-đun tiền xử lý:** Phần này được dành để xử lý trước các bản đồ đặc trưng một cách riêng biệt và mô-đun này thường gồm một lớp tích chập được liên kết với mỗi tỷ lệ hình ảnh. Cấu hình tối ưu bao gồm một lớp tích chập có kích thước 5×5 duy nhất trên mỗi quy mô.

– **Mô-đun tổng hợp:** Dùng để tổng hợp các RoI đơn lẻ thành một RoI duy nhất. Việc tổng hợp các RoI có thể giảm thiểu số lượng các đặc trưng cần phải tính cho lớp tiếp theo và điều này giúp mạng tập trung vào quá trình đào tạo từ đó giúp mạng ổn định hơn.

– **Mô-đun hậu xử lý:** Hậu xử lý là một bước xây dựng bổ sung được áp dụng cho các đặc trưng đã hợp nhất trước đó. Nó cho phép mạng tìm hiểu các đặc trưng toàn cục, xem xét trên tất cả các quy mô. Mục tiêu chính của lớp này là loại bỏ thông tin vô ích.



Hình 5. Kiến trúc kết hợp GRoIE và Double-Head (GRoIE Double-Head)

3.3. Đề xuất kết hợp GRoIE và Double-Head

Như đã đề cập tại 3.1 và 3.2, Double-Head và GRoIE là các loại kiến trúc phổ biến được thêm vào để cải thiện kết quả của mô hình. Do đó, những kiến trúc này có thể được coi là một mô-đun có thể dễ dàng kết hợp với các phương pháp phát hiện đối tượng khác. Đó là ý tưởng chính để nghiên cứu này xây dựng một kiến trúc mới áp dụng cả hai phương pháp GRoIE và Double-Head. Việc áp dụng GRoIE giúp trích xuất và kết hợp các đặc trưng RoI từ các tầng FPN, giúp đặc trưng mang thông tin ngữ nghĩa mạnh mẽ, tăng hiệu quả dự đoán tọa độ hộp giới hạn và phân lớp. Bên cạnh đó, Double-Head khác với các phương pháp 02 giai đoạn truyền thống. Thay vì sử dụng hai lớp kết nối đầy đủ để dự đoán cùng lúc tọa độ và lớp đối tượng, thì Double-Head chia ra hai nhánh để giải quyết công việc này. Việc phân lớp đối tượng sẽ do hai lớp kết nối đầy đủ đảm trách, còn hồi quy tọa độ sẽ được dự đoán bởi hai lớp tích chập. Việc hồi quy tọa độ bằng cách sử dụng các lớp tích chập đã được chứng minh là tốt hơn so với hai lớp kết nối đầy đủ. Trong mô hình Double-Head, hai lớp tích chập được sử dụng để hồi quy tọa độ có kích thước đầu ra là $7 \times 7 \times 1024$, kích thước đầu ra này được dùng mặc định như trong nghiên cứu. Với các nhận định trên, phương pháp kết hợp Double-Head với GroIE được tiến hành với tên gọi là GRoIE Double-Head, và đánh giá hiệu suất phát hiện phương tiện giao thông từ không ảnh của kiến trúc mới này. Kiến trúc mô hình cụ thể được trực quan ở Hình 5.

4. KẾT QUẢ THỰC NGHIỆM

4.1. Bộ dữ liệu AERIAU

Nghiên cứu được thực hiện trên bộ dữ liệu AERIAU (Chung et al., 2020). Đây một bộ dữ liệu từ trên xuống được chụp bởi máy bay không người lái dựa trên ba bộ dữ liệu video công cộng, cụ thể là VisDrone2018 (Zhu et al., 2018), KIT AIS (Weisbrich et al., 2012) và Aerial Open Source (Dertat et al., 2018). Bộ dữ liệu mới chứa 1.474 hình ảnh và 56.609 hộp đối tượng. Sau đó, AERIAU áp dụng các phương pháp tăng cường dữ liệu khác nhau bao gồm cắt xén và xoay ngẫu nhiên hình ảnh hiện có. Các hình ảnh trong bộ dữ liệu sử dụng bối cảnh, kích thước, góc nhìn và độ cao không gian khác nhau, độ cao không gian được sử dụng trong bộ dữ liệu từ 55 m đến trên 80 m. Điều này làm tăng sự đa dạng của các đối tượng trong bộ dữ liệu. Bộ dữ liệu không ảnh AERIAU sau khi áp dụng hai kỹ thuật tăng cường dữ liệu (AERIAU + Cắt xén + Quay ngẫu nhiên hình ảnh) bao gồm 7.792 ảnh định dạng JPG 4 nhãn tương ứng và 131.119 đối tượng được gán nhãn. Trong đó, tập huấn luyện (Train) 6.086 ảnh, tập thẩm định (Validation) 1.522 ảnh, tập kiểm tra (Test) 182 ảnh. Số lượng cụ thể các nhãn đối tượng trong bộ dữ liệu được trình bày trong Bảng 1.

Bảng 1. Thống kê chi tiết số lượng nhãn đối tượng trong bộ dữ liệu AERIAU

Object	Train	Validation	Test	Total
Car	84.253	20.999	3.704	108.956
Truck	5.260	1.349	77	6.686
Bus	2.607	625	264	3.496
Motor	8.662	2.262	1.057	11.981

(Chung et al., 2020)

4.2. Cấu hình thực nghiệm và chỉ số đánh giá

Toàn bộ quá trình thực nghiệm được triển khai trên môi trường Google Colab Pro. Nghiên cứu này tiến hành huấn luyện Double-Head trên MMDetection framework V2.10.0 (Chen et al., 2019) sử dụng cấu hình mặc định với kiến trúc

xương sống ResNet50 và ResNeXt-101 được huấn luyện trong vòng 12 epochs. SSD được áp dụng làm hàm kích hoạt với weight decay là 1e-4, learning rate được khởi tạo với giá trị 0,001 và động lượng (momentum) là 0,9.

Sau giai đoạn thực nghiệm là quá trình đánh giá mô hình Double-Head trên bộ dữ liệu AERIAU, độ đo mAP (Mean Average Precision) (Lin et al., 2014) được sử dụng để đánh giá trong nghiên cứu. Kết quả có được trên các độ đo AP_{50} và AP_{75} tương ứng với ngưỡng IoU (Intersection over Union) lần lượt là 0,5 và 0,75. Lần lượt tính AP (Average Precision) của mỗi lớp xuất hiện trong trang với các ngưỡng IoU khác nhau, từ đó tính trung bình để lấy AP của lớp đó. Sau khi có AP của tất cả các lớp, kết quả cuối cùng cho mô hình sẽ được đưa ra bằng việc tính trung bình để có được độ đo mAP.

4.3. Kết quả và thảo luận

Báo cáo kết quả thực nghiệm trên bộ dữ liệu AERIAU (Chung et al., 2020) được thống kê tại Bảng 2. Double-Head sử dụng kiến trúc xương sống ResNeXt-101 (37,09% mAP) là vượt trội hơn so với ResNet-50 (35,36% mAP). Double-Head ResNeXt-101 cho kết quả thấp hơn ở các lớp "Ô tô" (57,1% so với 58,9%), "Xe tải" (41,8% so với 54,0%), "Mô tô" (1,7% so với 3,9%) nhưng cao vượt trội ở lớp "Xe buýt" (47,9% so với 21,8%) so với Double-Head sử dụng ResNet-50. Điều này cho thấy một kiến trúc mạng sâu như ResNeXt-101 phát hiện tốt các đối tượng có kích thước lớn, nhưng lại cho hiệu quả không tốt trên các đối tượng nhỏ. Do ResNeXt-101 cho kết quả mAP tốt nhất nên nó được dùng trên phương pháp kết hợp GRoIE Double-Head đề xuất. Kết quả AP_{50}, AP_{75}, mAP lần lượt là 59,4%, 48,7%, 38,0%, cao hơn 1,19% 4,88% 0,91% so với Double Head nguyên bản sử dụng ResNet-50. GRoIE Double-Head cho kết quả vượt trội trên lớp "Xe buýt" so với Double-Head + ResNeXt-101, trong khi đó vẫn giữ được hiệu quả phát hiện tương đối với lớp "Ô tô". Tuy nhiên, ở lớp "Xe tải" và "Mô tô" thì mô hình kết hợp cho hiệu suất kém hơn Double-Head nguyên bản. Tuy nhiên, với độ đo mAP được giá trị cao nhất thì phương pháp GRoIE đề xuất vẫn hiệu quả hơn so với các thử nghiệm được báo cáo ở Bảng 2.

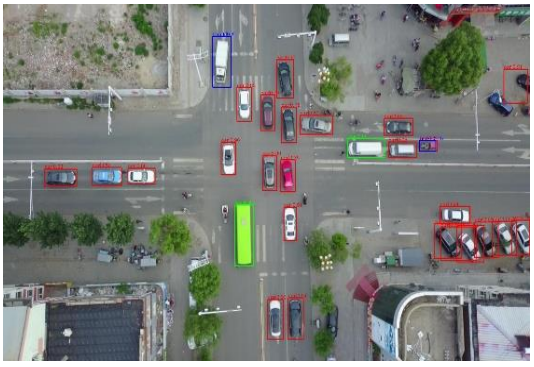
Bảng 2. So sánh kết quả thực nghiệm các phương pháp trên bộ dữ liệu AERIAU (%)

Method	Backbone	Car	Bus	Truck	Motor	AP_{50}	AP_{75}	mAP
YOLOv3	Darknet-53	54,56	12,21	**62,72**	**10,86**	-	-	35,08
Double-Head	ResNet-50	**58,9**	21,8	54,0	3,9	54,08	43,1	35,36
Double-Head	ResNeXt-101	57,1	47,9	41,8	1,7	58,21	43,82	37,09
GRoIE Double-Head	ResNeXt-101	55,4	**54,5**	39,5	2,6	**59,4**	**48,7**	**38,0**

Phương pháp một giai đoạn (YOLOv3) đã được Chung et al. (2020) thực nghiệm và đánh giá trên bộ dữ liệu AERIAU, do đó kết quả này được so sánh với kết quả của phương pháp đề xuất. Kết quả cho thấy Double-Head phát hiện đối tượng "Ô tô" và "Xe buýt" ở bộ dữ liệu AERIAU tốt hơn YOLOv3, trong khi các đối tượng thuộc lớp "Mô tô" và "Xe tải" lại cho kết quả thấp hơn.

Điều này cho thấy phương pháp 01 giai đoạn như YOLOv3 có hiệu quả phát hiện tốt hơn trên các đối tượng nhỏ so với các phương pháp 02 giai đoạn. YOLOv3 được chứng minh tốt hơn trên các đối tượng nhỏ, (Liu et al., 2016) đã so sánh hiệu suất phát hiện đối tượng nhỏ, lớn và trung bình của các phương pháp phát hiện một giai đoạn và hai giai đoạn dựa trên tập dữ liệu MS COCO. Trong đó, YOLOv3 đạt điểm số cao thứ ba đối với các đối tượng nhỏ (18,3% APs), chỉ sau Retinane (21,8% APs với kiến trúc xương sống ResNet-101 và 24,1% APs với kiến trúc xương sống ResNeXt-101). YOLOv3 phát hiện đối tượng nhỏ tốt hơn bởi vì bản chất YOLOv3 chia ảnh theo từng lưới và phát hiện dựa trên các lưới này. Bên cạnh đó, YOLOv3 sử dụng 3 tỷ lệ để chia các lưới trên ảnh, càng nhiều lưới thì mô hình sẽ nhìn được rõ hơn các đối tượng nhỏ.

a. Dự đoán sai và bao không hết đối tượng



b. Chồng chéo các hộp giới hạn

c. Hộp giới hạn không chứa đối tượng

d. Bỏ sót đối tượng

Hình 6. Trực quan những hình ảnh dự đoán chưa tốt sau khi thực nghiệm trên bộ dữ liệu AERIAU

Mỗi màu sắc của các hộp giới hạn thể hiện cho đối tượng khác nhau, đối tượng màu đỏ đại diện cho các đối tượng xe ô tô, màu xanh lam cho xe tải, màu xanh lục cho xe buýt và màu vàng cho xe mô tô.

Phương pháp Double-Head mang lại AP khá tốt, cải thiện tỷ lệ phát hiện đối tượng lên khoảng 2% so với thực nghiệm gốc. Tuy nhiên, độ chính xác giữa các lớp là không nhất quán. Nhìn bảng 2 ta thấy, độ tự tin cao nhất thuộc về lớp "Ô tô", mô hình phân loại khá chính xác đối tượng trên với kết quả 57,1% AP. Điều này là dễ hiểu vì lớp đối tượng "Ô tô" có nhiều nhãn đối tượng huấn luyện nhất, tỷ lệ kích thước của hầu hết các đối tượng này so với hình ảnh trong bộ dữ liệu cũng đã giúp mô hình trích xuất các đặc trưng rõ ràng hơn. Tuy AP khá cao nhưng việc dự đoán đối tượng "Ô tô" vẫn gặp một vài lỗi như bỏ sót đối tượng (Hình 6 (d)) và chồng chéo các hộp giới hạn (Hình 6 (b)). Kết quả thực nghiệm của lớp đối tượng "Mô tô" là 1,7% AP, một kết quả rất thấp do lớp "Mô tô" là lớp đối tượng có kích thước rất nhỏ. Phát hiện đối tượng nhỏ đã và đang là thách thức lớn nhận được nhiều sự quan tâm trong cộng đồng thị giác máy tính. Kết quả thực nghiệm trên lớp "Xe buýt" (47,9% AP) và lớp "Xe tải" (41,8% AP) đạt hiệu suất phát hiện ở mức tương đối và vẫn có những sai sót đáng quan tâm như việc bỏ sót đối tượng và nhầm lẫn với các đối tượng khác (Hình 6 (a)), các hộp giới hạn bao không hoàn toàn đối tượng (Hình 6 (a)). Đánh giá mô hình thực nghiệm dựa trên những phân tích trên, mô hình Double-Head đạt kết quả phát hiện phương tiện giao thông tương đối khả quan, mô hình đạt được hiệu suất cao với các đối tượng "Ô tô" và "Xe buýt". Tuy nhiên, vẫn xảy ra một số vấn đề như: bỏ sót và nhầm lẫn các đối tượng xuất hiện rõ ràng trong hình ảnh (Hình 6 (a)), Hình 6 (d)); xảy ra tình trạng chồng chéo các hộp giới hạn (Hình 6 (b)); bao không hết đối tượng (Hình 6 (a)); hộp giới hạn không chứa đối tượng (Hình 6 (c)). Đánh giá kiến trúc Double-Head kết hợp GRoIE (hay còn gọi GRoIE Double-Head) được đề xuất trong bài báo này, mô hình đã đạt được kỳ vọng khi kết quả trên lớp đối tượng "Xe buýt" là cao nhất trong tất cả thực nghiệm được tiến hành (54,5% AP), kết quả trên lớp đối tượng "Ô tô" cũng rất cao (55,4% AP). Tuy nhiên, mô hình không hoạt động tốt như mong đợi trên các lớp đối tượng nhỏ và vừa khi lớp "Xe tải" chỉ đạt được 39,5% AP (thấp nhất) và 2,6% AP trên lớp "Mô tô".

5. KẾT LUẬN

Phương pháp áp dụng cấu trúc hai đầu – Double-Head được thực nghiệm và đánh giá để phát hiện

phương tiện giao thông từ không ảnh trên bộ dữ liệu AERIAU. Kết quả cho thấy rằng so với phương pháp phát hiện một giai đoạn (YOLOv3), Double-Head phát hiện các lớp đối tượng "Ô tô" và "Xe buýt" dễ dàng và hiệu quả hơn. Ngược lại, đối tượng "Xe tải" và "Mô tô" cho hiệu suất thấp hơn. Ngoài ra, mô hình GRoIE Double-Head của nghiên cứu này hoạt động tốt đối với các đối tượng có kích lớn và trung bình, cải thiện được hiệu suất phát hiện. Trong khi đó, mô hình cho kết quả không khả quan khi dự đoán các đối tượng có kích thước nhỏ. Phát hiện này có thể dẫn đến các nghiên cứu sâu hơn như thay đổi cấu hình của từng tùy chọn để đạt được cài đặt phù hợp nhất với tất cả các lớp đối tượng được xem xét. Các vấn đề đã được trình bày trong nghiên cứu góp phần mang lại nguồn cảm hứng và đóng góp hữu ích cho các nghiên cứu liên quan.

LỜI CẢM TẠ

Nghiên cứu được thực hiện tại Phòng thí nghiệm Truyền thông Đa phương tiện (MMLab), Trường Đại học Công nghệ Thông tin, Đại học Quốc gia Thành phố Hồ Chí Minh.

TÀI LIỆU THAM KHẢO

Bae, W., Noh, J., & Kim, G. (2020). Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision* (pp. 618-634). Springer, Cham. https://doi.org/10.1007/978-3-030-58555-6_37

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, X., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, O., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, D., Wang, J., Shi, J., Ouyang, W., Loy, C. C., & Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

Chung, Q. M., Le, T. D., Dang, T. V., Vo, N. D., Nguyen, T. V., & Nguyen, K. (2020). Data augmentation analysis in vehicle detection from aerial videos. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)* (pp. 1-3). IEEE. https://doi.org/10.1007/978-3-030-58555-6_37

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems, 29*.

Dertat, A. (2018). Applied deep learning-part 1: Artificial neural networks, 2017. *URl:* https://towardsdatascience.com/applied-deep-learningpart-1-artificial-neural-networks-d7834f67a4f6.

Tổng cục Đường bộ Việt Nam. (2021). *Ô nhiễm môi trường giao thông tại VN: Thực trạng và giải pháp*. https://drvn.gov.vn/tin-tuc/tin-tuc-su-kien/o-nhiem-moi-truong-giao-thong-tai-vn-thuc-trang-va-giai-phap2.html?site=20830

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969). https://doi.org/10.1109/ICCV.2017.322

Ho, N., Pham, M., Vo, N. D., & Nguyen, K. (2020). Vehicle detection at night time. In *2020 7th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 250-255). IEEE. https://doi.org/10.1109/NICS51282.2020.9335870

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980-2988). https://doi.org/10.1109/ICCV.2017.324

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision* (pp. 21-37). Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

Liu, Z., Zhang, W., Gao, X., Meng, H., Tan, X., Zhu, X., Xue, Z., Ye, X., Zhang, H., Wen, S., & Ding, E. (2020). Robust movement-specific vehicle counting at crowded intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 614-615). https://doi.org/10.1109/CVPRW50498.2020.00315

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788). https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271). https://doi.org/10.1109/CVPR.2017.690

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

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

Rossi, L., Karimi, A., & Prati, A. (2021). A novel region of interest extraction layer for instance segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 2203-2209). https://doi.org/10.1109/ICPR48806.2021.9412258

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). https://doi.org/10.1109/CVPR.2015.7298594

Weisbrich, W. I. (2012). Kit-ipf-forschung – downloads. http://www.ipf.kit.edu/downloads.php

Zhu, P., Wen, L., Bian, X., Ling, H., & Hu, Q. (2018). Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*.