Nghiên cứu bài toán phát hiện và phân loại phương tiện giao thông từ tín hiệu hình ảnh

■ **TS. LÊ QUYẾT TIẾN; ThS. TRẦN ĐÌNH VƯƠNG**
Trường Đại học Hàng hải Việt Nam

TÓM TẮT: Đóng góp chính của bài báo là đưa ra giải pháp cho bài toán phát hiện và phân loại phương tiện giao thông từ tín hiệu hình ảnh, đồng thời đề xuất giải pháp nâng cao hiệu quả cho bài toán. Hướng tiếp cận bài toán dựa trên phương pháp học sâu sử dụng mô hình YOLO, một trong các mô hình điển hình nhất hiện nay với bài toán phát hiện và phân loại đối tượng. Giải pháp đề xuất được nghiên cứu và đánh giá một cách chặt chẽ và cho thấy hiệu quả cải thiện rõ rệt nhờ sử dụng phương pháp tăng dữ liệu. Tăng dữ liệu ngoài giúp việc gia tăng số lượng dữ liệu tránh hiện tượng quá khớp dữ liệu trong học sâu còn đa dạng hóa các trường hợp giúp cho hệ thống khi triển khai có thể thích ứng được với nhiều điều kiện hoạt động khác nhau. Kết quả thực nghiệm cho thấy phương pháp đề xuất có độ chính xác trung bình (mAP) đạt 0,732.

TỪ KHÓA: Phát hiện đối tượng, phân loại hình ảnh, phát hiện phương tiện, học sâu.

ABSTRACT: The main contributions of this paper is to propose a solution solving the task of vehicle detection and classification and to study a data augmentation solution for improving the accuracy of the task. The considered approach is based on deep-learning methods using the YOLO model, one of the most typical models for the task of object detection and classification. The proposed solutions have been studied and evaluated strictly and the experimental results show a significant improvement in the performance. Data augmentation does not only help increasing the number of samples to prevent over-fitting situations but also help implemented systems adapting with different contexts. Experimental results show that the proposed method has the mAP at 0.732.

KEYWORDS: Object detection, image classification, vehicle detection, deep learning.

1. ĐẶT VẤN ĐỀ

TNGT là một trong các vấn đề tồn đọng trong xã hội ngày nay mặc dù nhiều giải pháp khác nhau đã được đưa ra bởi các bộ, ban, ngành. Tới nay, các thiệt hại để lại từ TNGT vẫn là những con số đáng báo động. Theo các báo cáo thống kê từ Cục CSGT, trong nửa đầu của năm 2021 đã có gần 6.300 vụ tai nạn xảy ra trên cả nước, làm bị thương gần 4.500 người và thiệt mạng hơn 3.000 người. Những con số trên thể hiện được mức độ phức tạp của tình hình giao thông tại Việt Nam. Những giải pháp khoa học công nghệ đã và đang là một hướng giải quyết hiệu quả để giải quyết bài toán ATGT.

Những năm gần đây, nhờ thành tựu trong khoa học, các giải pháp công nghệ đang tạo ra những thay đổi lớn trong và ngoài nước. Các camera hành trình gắn kèm trên phương tiện giao thông đang trở nên ngày càng phổ biến. Mặc dù vậy, hiện tại ở Việt Nam, camera hành trình chủ yếu được sử dụng chỉ để ghi lại hình ảnh trong quá trình tham gia giao thông mà chưa được khai thác trong tác vụ trợ giúp người điều khiển phương tiện. Phát hiện và phân loại các phương tiện giao thông có thể coi là một trong các chìa khóa mở ra giải pháp giảm thiểu rủi ro tham gia giao thông. Hiện tại, bài toán này đang tồn tại nhiều khó khăn. Thứ nhất, các camera có chất lượng rất đa dạng cũng như điều kiện ánh sáng, môi trường tại các khu vực giao thông cũng rất khác nhau dẫn tới một số hệ thống phát hiện và phân loại phương tiện giao thông chỉ hoạt động tốt trên một kiểu môi trường nhất định hoặc một nhóm loại thiết bị nhất định. Thứ hai, bộ dữ liệu cho bài toán này còn chưa thật sự phong phú dẫn đến tình trạng dữ liệu không đủ để huấn luyện các mô hình học sâu vốn cần lượng lớn dữ liệu. Trong nghiên cứu này, bài toán phát hiện và phân loại các phương tiện giao thông được tập trung giải quyết. Giải pháp áp dụng các kỹ thuật tăng dữ liệu nhằm đa dạng hóa dữ liệu và gia tăng số lượng mẫu giúp mô hình có thể hoạt động linh hoạt và hiệu quả hơn.

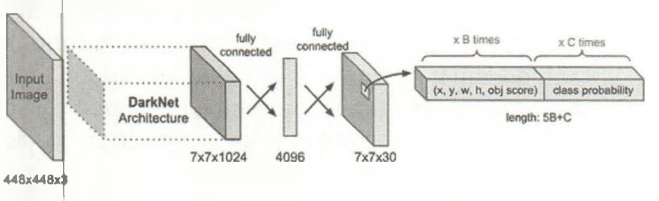
2. BỐI CẢNH NGHIÊN CỨU

Phát hiện và phân loại phương tiện giao thông nằm trong chuỗi các bài toán phát hiện và phân loại đối tượng (nói ngắn gọn là phát hiện đối tượng). Bài toán phát hiện đối tượng bao gồm định vị và phân loại đối tượng. Trong đó, bài toán định vị đối tượng xác định hộp giới hạn bao quanh từng đối tượng, còn bài toán phân loại đối tượng phân lớp (gán nhãn) của đối tượng trong hộp giới hạn.

Bài toán phát hiện đối tượng có nhiều hướng giải quyết khác nhau. Ở giai đoạn trước khi các mô hình học sâu được đưa ra, các hướng giải quyết như phương pháp Viola & Jones [1, 17], sử dụng biểu đồ thống kê định hướng

gradient [2], bộ phát hiện phần biến dạng [3] được kết hợp với cửa sổ trượt (sliding window). Sự phát triển của các mô hình học sâu đã thay đổi hướng tiếp cận của bài toán sang các mô hình mạng nơ-ron tích chập (Convolutional Neural Network - CNN) [8]. Hai chuỗi mô hình điển hình cho hướng tiếp cận này là R-CNN (Regions with Convolutional Neural Network - mạng nơ-ron tích chập vùng) [4, 6, 7] và YOLO (You Only Look Once - bạn chỉ nhìn một lần) [9, 10, 11, 12, 22]. Tổng quan, các đối tượng được họ mô hình R-CNN phát hiện qua giai đoạn một: đề xuất vùng và gian đoạn hai: phân loại vùng. Với một hướng tiếp cận khác, chuỗi mô hình YOLO thực hiện đề xuất và phân loại vùng chỉ trong một giai đoạn. Nhờ đó, các mô hình YOLO có thời gian thực thi nhanh hơn nhiều R-CNN nhưng vẫn cho độ chính xác tốt.

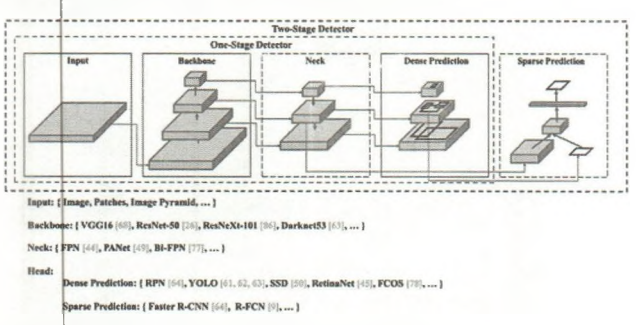
YOLOV1 [9] được đề xuất đầu tiên trong họ mô hình. Mô hình này chia hình ảnh thành nhiều ô (cell), mỗi ô được phân ra chịu trách nhiệm đưa ra khung giới hạn và phân lớp của khung có tâm thuộc vùng của ô đó. Một mạng học sâu (ví dụ LeNet, AlexNet [5], VGG [13], GoogLeNet [14], ResNet [15], EfficientNet [16]) được sử dụng để tính toán các bản đồ đặc trưng. Các đặc trưng được kết nối với các lớp kết nối đầy đủ để đưa ra phân lớp, vị trí và kích thước của các khung đối tượng *(Hình 2.1)*.



Hình 2.1: Kiến trúc YOLOV1

YOLOV2 [10] được giới thiệu sau đó với sự nâng cấp là đưa vào các lớp chuẩn hóa (normalization layers) và thay thế các lớp kết nối đầy đủ bởi các lớp hộp neo (anchor box layer) để khởi tạo hộp giới hạn trước và điều chỉnh lại vị trí và kích thước ở các bước tiếp theo. YOLOV3 [11] có một số sự thay đổi của mạng tích chập và phát hiện đối tượng trên nhiều tỷ lệ ảnh khác nhau thay vì chỉ một tỷ lệ như các mô hình trước. Nhờ đó, YOLOV3 có thể phát hiện được các đối tượng ở các kích thước đa dạng hơn.

YOLOV4 [12] được nâng cấp kiến trúc gồm xương sống (backbone), cổ (neck) và đầu (head) như *Hình 2.2*. YOLOV5 [22] có kiến trúc tương tự nhưng sử dụng CSPDarknet làm xương sống, PANet làm cổ và dùng lớp Yolo cho phần đầu.



Hình 2.2: Kiến trúc YOLOV4

Bài toán phát hiện và phân loại phương tiện giao thông đã và vẫn đang được đưa ra nhiều trong các nghiên cứu. Các hướng nghiên cứu hiện tại chủ yếu tập trung vào hướng tiếp cận học sâu [18, 19, 20, 21]. Các nghiên cứu trước đây thường sử dụng trên một bộ dữ liệu xác định và mô hình được huấn luyện với bộ dữ liệu nào thường sẽ chỉ hoạt động ổn định với với môi trường tương tự. Thực tế triển khai cho thấy các camera có chất lượng rất khác nhau cùng sự đa dạng môi trường dẫn tới sự không ổn định và kém linh hoạt ở các hệ thống thực tế. Trong bài báo này, bài toán phát hiện và phân loại phương tiện giao thông sẽ được tập trung giải quyết, đồng thời giải pháp nâng cao hiệu năng bài toán cũng được đưa ra nghiên cứu và đánh giá.

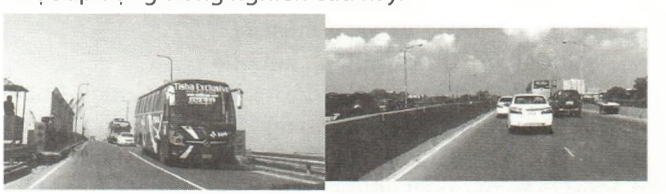
3. BÀI TOÁN PHÁT HIỆN VÀ PHÂN LOẠI PHƯƠNG TIỆN GIAO THÔNG

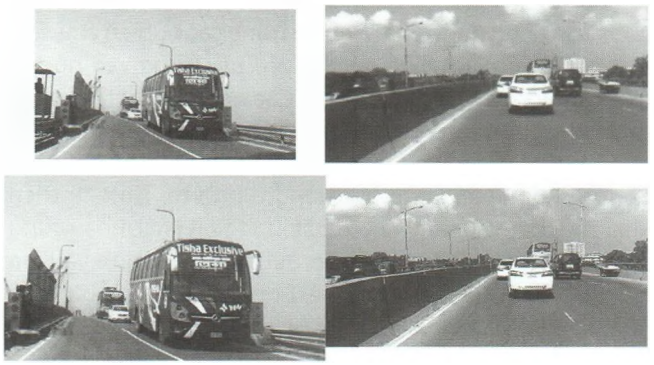
3.1. Phương pháp nghiên cứu

Xuất phát từ ý tưởng sử dụng phương pháp tăng dữ liệu để cải thiện hiệu năng bài toán, nghiên cứu được thực hiện theo hai hướng: huấn luyện mô hình trên bộ dữ liệu gốc (không được tăng dữ liệu) và huấn luyện mô hình trên bộ dữ liệu được tăng cường. Kết quả đánh giá của hai hướng được so sánh để đánh giá xem việc áp dụng phương pháp tăng dữ liệu cải thiện hiệu suất bài toán đến đâu. Nghiên cứu này hướng tới mục tiêu chạy thời gian thực nên nhóm tác giả lựa chọn mô hình YOLOV5 (vì sự cân bằng về độ chính xác và thời gian thực thi) trong nghiên cứu này.

Huấn luyện một mô hình học sâu đòi hỏi một lượng dữ liệu lớn để tránh tình trạng mô mình bị mất đi tính tổng quát (over-fitting). Đây là một thách thức không nhỏ do việc thu thập dữ liệu chưa bao giờ dễ dàng. Ngoài ra, dữ liệu thu thập khó đa dạng như các dữ liệu thực tế do chất lượng đa dạng của các camera thực tế ảnh hưởng tới chất lượng hình ảnh thu nhận. Ngoài ra, sự đa dạng của điều kiện ánh sáng, môi trường, khung cảnh cũng góp phần tạo ra những sai lệch. Do đó, nhóm tác giả hướng tới giải pháp tăng dữ liệu để đa dạng hóa và tăng số lượng dữ liệu huấn luyện giúp mô hình học sâu được huấn luyện hiệu quả hơn.

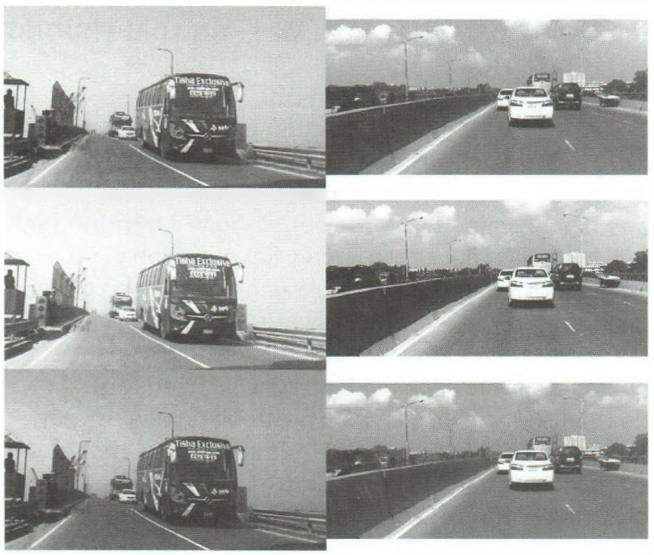
Chất lượng đa dạng của các camera là nguyên nhân nhóm tác giả cân nhắc việc đa dạng hóa độ phân giải *(Hình 3.1a)* và độ nét *(Hình 3.1b)*. Ngoài ra, điều kiện ánh sáng môi trường cũng là một yếu tố phức tạp nên việc thay đổi điều kiện ánh sáng *(Hình 3.2a)* và độ tương phản *(Hình 3.2b)* cũng được thực hiện trong phương pháp tăng dữ liệu được đề xuất. Những biến dạng không gian có thể xảy ra do góc thu nhận ảnh nên việc thay đổi tỷ lệ khung hình (tỷ lệ chiều rộng/chiều cao - *Hình 3.3a*) và góc quan sát (*Hình 3.3b*) cũng được cân nhắc. Thêm vào đó, một biện pháp tăng dữ liệu thường được sử dụng là lật ảnh (image fliptation) cũng được áp dụng trong nghiên cứu này.





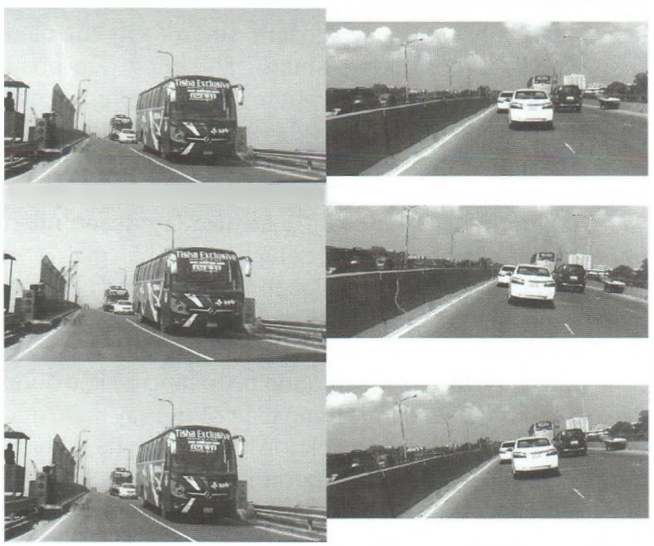
a) b)

Hình 3.1: Tăng dữ liệu nhờ thay đổi độ phân giải (a) và độ nét (b)



a) b)

Hình 3.2: Tăng dữ liệu nhờ thay đổi độ sáng (a) và độ tương phản (b)



a) b)

Hình 3.3: Tăng dữ liệu nhờ thay đổi tỷ lệ khung hình (a) và góc quan sát (b)

3.2. Cài đặt, thực nghiệm, kết quả

Nghiên cứu được cài đặt trên môi trường Google Colab (GPU Nvidia K80) sử dụng ngôn ngữ lập trình Python.

Thí nghiệm được thực hiện trên bộ dữ liệu nhóm tác giả thu thập từ nhiều nguồn (chủ yếu khai thác từ nguồn Kaggle) với hơn 3.000 hình ảnh chụp đường phố kèm chứa nhãn 11 loại phương tiện giao thông khác nhau bao gồm: xe buýt, xe kéo, xe máy, ô tô con, xe ba bánh, xe chở hàng, xe thể thao/đa dụng, xe khách nhỏ, xe tải, xe đạp, xe quân đội. Tất cả phương tiện trong hình đều được khoanh vùng và gán nhãn. Bộ dữ liệu được chia ngẫu nhiên thành tập huấn luyện chứa 2.455 hình ảnh (19.402 đối tượng được gán nhãn) và tập đánh giá chứa 547 hình ảnh (3.476 đối tượng được gán nhãn).

Dựa trên bộ dữ liệu này, hai thí nghiệm khác nhau sẽ được thực hiện. Thí nghiệm thứ nhất thực hiện huấn luyện mô hình YOLOV5 sử dụng 2.455 hình ảnh trong tập huấn luyện. Thí nghiệm thứ hai sử dụng 2.445x3 (7.335) hình ảnh để huấn luyện mô hình YOLOV5 (ngoài phiên bản gốc của mỗi hình ảnh *(Hình 3.4a)* sẽ có thêm hai phiên bản tăng dữ liệu được sinh ra dựa trên các phép biến đổi thay đổi độ phân giải, độ nét, độ sáng, độ tương phản, tỷ lệ khung hình, góc quan sát và lật ảnh, các tham số biến đổi được chọn ngẫu nhiên và không làm sai khác quá 15% so với ảnh gốc. Một phiên bản được sinh ra theo hướng gia tăng các chỉ số *(Hình 3.4b)* và phiên bản còn lại được sinh ra theo hướng giảm các chỉ số *(Hình 3.4c)*. Trong cả hai thí nghiệm, các hình ảnh được chuẩn hóa về kích thước 640x640 bằng cách thay đổi kích thước (resize) và thêm vùng đệm (padding). Cả hai mô hình được huấn luyện trong 50 vòng và dừng sớm nếu bị hiện tượng khớp quá mức dữ liệu (over-fitting). Việc đánh giá mô hình trong cả hai thí nghiệm được thực hiện trên 547 hình ảnh thuộc tập đánh giá dựa trên chỉ số gợi nhớ (recall), độ chính xác (precision) và giá trị chính xác trung bình (mean average precision - mAP).

Hình 3.4: Các phiên bản tăng cường dữ liệu

Bảng 3.1. Kết quả đánh giá hai mô hình trong hai thí nghiệm

Mô hình không sử dụng phương pháp tăng dữ liệu	
Độ chính xác (precision)	0,574
Chỉ số gợi nhớ (recall)	0,519
Giá trị chính xác trung bình (mAP)	0,513
Mô hình sử dụng phương pháp tăng dữ liệu	
Độ chính xác (precision)	0,669
Chỉ số gợi nhớ (recall)	0,720
Giá trị chính xác trung bình (mAP)	0,732

Kết quả của thí nghiệm nêu trên được thể hiện trong *Bảng 3.1*. Hiệu suất ở thí nghiệm thứ nhất không thật sự ấn tượng với độ chính xác 0,574, chỉ số gợi nhớ 0,519 và độ chính xác trung bình 0,513. Việc hiệu suất của mô hình chỉ dừng ở mức này có thể giải thích do lượng dữ liệu trong bộ huấn luyện còn chưa nhiều. Thêm vào đó, trong bộ dữ liệu của bài toán bao gồm nhiều hình ảnh được chụp dưới nhiều điều kiện khác nhau bởi nhiều thiết bị khác nhau. Điều này dẫn đến có khả năng dữ liệu trong tập huấn luyện và tập đánh giá không tương đồng nên hiệu suất bị giảm. Nói cách khác, bộ dữ liệu được sử dụng mặc dù số lượng ảnh không quá nhiều cộng thêm việc tập huấn luyện không có được sự đa dạng nên có thể coi là một bộ dữ liệu chưa phong phú. Ở thí nghiệm thứ hai, việc sử dụng phương pháp tăng dữ liệu đem lại hiệu quả tốt hơn *(Hình 3.5)*. Độ chính xác, chỉ số gợi nhớ và giá trị chính xác trung bình của mô hình huấn luyện trên tập dữ liệu gia tăng đều cao hơn so với các giá trị lần lượt 0,669, 0,720 và 0,732. Có thể thấy, việc tăng dữ liệu giúp việc huấn luyện trở nên hiệu quả hơn. Mô hình nhờ đó hoạt động linh hoạt và thích ứng được với nhiều hoàn cảnh khác nhau hơn. Nói cách khác, bài toán phát hiện và phân loại phương tiện giao thông có thể cải thiện hiệu suất một cách rõ rệt nhờ việc áp dụng phương pháp tăng dữ liệu được đề xuất.

Hình 3.5: Kết quả chạy thử nghiệm hệ thống

4. KẾT LUẬN

Với mục tiêu xây dựng và nâng cao hiệu suất hệ thống phát hiện và phân loại phương tiện giao thông, nghiên cứu đã đề xuất hai hướng tiếp cận cho bài toán: hướng tiếp cận thông thường sử dụng bộ dữ liệu không áp dụng phương pháp tăng dữ liệu và hướng tiếp cận sử dụng bộ dữ liệu áp dụng phương pháp tăng dữ liệu. Các hướng tiếp cận được nghiên cứu và đánh giá chặt chẽ trên bộ dữ liệu thu thập bởi nhóm tác giả với hơn 3.000 hình ảnh và gần 23.000 phương tiện giao thông được định vị và gán nhãn. Kết quả thực nghiệm cho thấy hướng sử dụng phương pháp tăng dữ liệu có hiệu quả tốt hơn việc không sử dụng phương pháp tăng dữ liệu với độ chính xác trung bình 0,732. Điều đó thể hiện được hiệu quả của giải pháp tăng cường hiệu quả cho bài toán được nhóm tác giả đề xuất.

Dựa trên các kết quả của nghiên cứu này, hướng nghiên cứu của nhóm tác giả trong tương lai là các hệ thống cảnh báo rủi ro khi tham gia giao thông và các hệ thống hỗ trợ lái xe tự động sử dụng cảm biến và camera.

Lời cảm ơn: Nghiên cứu này được tài trợ bởi Trường Đại học Hàng hải Việt Nam trong Đề tài mã số DT21-22.59.

Tài liệu tham khảo

[1]. Viola, Paul and Michael Jones (2001), *Rapid object detection using a boosted cascade of simple features*, CVPR, Vol. 1, doi: 10.1109/CVPR.2001.990517.

[2]. Dalal, N., Triggs, B. (2005), *Histograms of oriented gradients for human detection*, CVPR, doi: 10.1109/CVPR.2005.177.

[3]. Cho, Hyunggi, et al. (2012), *Real-time pedestrian detection with deformable part models*, IEEE Intelligent Vehicles Symposium, doi: 10.1109/IVS.2012.6232264.

[4]. Girshick, Ross, et al., *Rich feature hierarchies for accurate object detection and semantic segmentation*, CVPR. 2014. (pp. 580-587), doi: 10.1109/CVPR.2014.81.

[5]. A. Krizhevsky, I. Sutskever and G. E. Hinton (May 2017), *ImageNet classification with deep convolutional neural networks*, Commun. ACM, vol.60, no.6, pp.84-90, doi: 10.1145/3065386.

[6]. Girshick, Ross (2015), *Fast r-cnn*, Proceedings of the IEEE international conference on computer vision, doi: 10.1109/ICCV.2015.169.

[7]. Ren, Shaoqing, et al (2015), *Faster r-cnn: Towards real-time object detection with region proposal networks*, Advances in neural information processing systems 28, pp.91-99, doi: 10.1109/TPAMI.2016.2577031.

[8]. K. Simonyan and A. Zisserman (2015), *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556 [cs], Apr., Accessed: Apr. 22, 2021. [Online]. Available: *http://www.arxiv.org/abs/1409.1556*.

[9]. Redmon, Joseph, et al. (2016), *You only look once: Unified, real-time object detection*, CVPR. doi: 10.1109/CVPR.2016.91.

[10]. Redmon, Joseph and Ali Farhadi, *YOLO9000: better, faster, stronger*, CVPR. 2017. doi: 10.1109/CVPR.2017.690.

[11]. Redmon, Joseph and Ali Farhadi, *Yolov3: An incremental improvement*, arXiv preprint arXiv:1804.02767 (2018), Available at: *http://arxiv.org/abs/1804.02767*.

[12]. Bochkovskiy, Alexey, Chien-Yao Wang and Hong-Yuan Mark Liao (2020), *Yolov4: Optimal speed and accuracy of object detection*, arXiv preprint arXiv:2004.10934, Available at: *http://arxiv.org/abs/2004.10934*.

[13]. Simonyan, Karen and Andrew Zisserman (2014), *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, Available at: *http://www.arxiv.org/abs/1409.1556*.

[14]. Szegedy, Christian, et al., *Rethinking the inception architecture for computer vision*, CVPR. 2016. doi: 10.1109/CVPR.2016.308.

[15]. He, Kaiming, et al., *Deep residual learning for image recognition*, CVPR. 2016. doi: 10.1109/CVPR.2016.90.

[16]. M. Tan and Q. V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, arXiv:1905.11946 [cs, stat], Sep. 2020, Accessed: Apr. 22, 2021. [Online]. Available at: *http://www.arxiv.org/abs/1905.11946*.

[17]. Ince, Omer F., et al. (2014), *Child and adult classification using ratio of head and body heights in images*, International Journal of Computer and Communication Engineering 3.2, doi: 10.7763/IJCCE.2014.V3.304.

[18]. Sang, J., Wu, Z., Guo, P., Hu, H., Xiang, H., Zhang, Q. and Cai, B. (2018), *An improved YOLOv2 for vehicle detection*, Sensors, 18(12), p.4272.

[19]. Wang, H., Yu, Y., Cai, Y., Chen, X., Chen, L. and Liu, Q. (2019), *A comparative study of state-of-the-art deep learning algorithms for vehicle detection*, IEEE Intelligent Transportation Systems Magazine, 11(2), pp.82-95.

[20]. Hu X, Xu X, Xiao Y, Chen H, He S, Qin J, Heng PA. (2018 Oct 1), *SINet: A scale-insensitive convolutional neural network for fast vehicle detection*, IEEE transactions on intelligent transportation systems, 20(3):1010-9.

[21]. Song H, Liang H, Li H, Dai Z, Yun X. (2019 Dec), *Vision-based vehicle detection and counting system using deep learning in highway scenes*, European Transport Research Review, 11(1):1-6.

[22]. Xu, R., Lin, H., Lu, K., Cao, L. and Liu, Y. (2021), *A forest fire detection system based on ensemble learning*, Forests, 12(2), p.217.

Ngày nhận bài: 15/5/2022
Ngày chấp nhận đăng: 11/6/2022
Người phản biện: TS. Nguyễn Hữu Tuân
TS. Trần Thị Hương