DEVELOPING A DEEP LEARNING MODEL FOR DETECTING CAMOUFLAGED MILITARY OBJECTS BASED ON THE YOLOV9 MODEL

Pham Thu Huong
Institute of Information Technology - AMST

ARTICLE INFO

Received: 20/9/2024
Revised: 13/11/2024
Published: 14/11/2024

KEYWORDS

Deep learning
Computer vision
Military camouflage
Object detection
YOLO

ABSTRACT

Detecting camouflaged military objects is a significant challenge as they are typically designed to blend into their surroundings. This paper proposes an automated method for detecting military camouflage using deep learning techniques, specifically the YOLOv9 (You Only Look Once) model. YOLOv9 is one of the advanced object detection models, renowned for its real-time processing capabilities and high accuracy. The YOLOv9 model was trained and evaluated on a specialized dataset, including images containing camouflaged military objects in various contexts. Experimental results show that the YOLOv9 model achieves high performance in detecting camouflaged military objects, with superior accuracy compared to traditional methods. This paper not only demonstrates the feasibility of applying deep learning to camouflage detection but also opens up new directions for research and applications in this field.

XÂY DỰNG MÔ HÌNH HỌC SÂU PHÁT HIỆN ĐỐI TƯỢNG NGỤY TRANG QUÂN SỰ DỰA TRÊN MÔ HÌNH YOLOv9

Phạm Thu Hương
Viện Công nghệ thông tin - Viện Khoa học và Công nghệ quân sự

THÔNG TIN BÀI BÁO

Ngày nhận bài: 20/9/2024
Ngày hoàn thiện: 13/11/2024
Ngày đăng: 14/11/2024

TỪ KHÓA

Học sâu
Thị giác máy tính
Ngụy trang quân sự
Phát hiện đối tượng
YoLo

TÓM TẮT

Phát hiện đối tượng ngụy trang quân sự là một thách thức lớn do các đối tượng này thường được thiết kế để hòa lẫn vào môi trường xung quanh. Bài báo này đề xuất một phương pháp tự động phát hiện ngụy trang quân sự bằng cách sử dụng kỹ thuật học sâu, cụ thể là mô hình YOLOv9 (You Only Look Once). YOLOv9 là một trong những mô hình phát hiện đối tượng tiên tiến, nổi bật với khả năng xử lý thời gian thực và độ chính xác cao. Mô hình YOLOv9 được huấn luyện và đánh giá trên tập dữ liệu đặc thù, bao gồm các hình ảnh chứa đối tượng ngụy trang quân sự trong nhiều bối cảnh khác nhau. Kết quả thử nghiệm cho thấy mô hình YOLOv9 đạt được hiệu suất cao trong việc phát hiện đối tượng ngụy trang quân sự, với độ chính xác vượt trội so với các phương pháp truyền thống. Bài báo này không chỉ chứng minh tính khả thi của việc áp dụng học sâu vào phát hiện ngụy trang mà còn mở ra những hướng đi mới cho các nghiên cứu và ứng dụng trong lĩnh vực này.

DOI: https://doi.org/10.34238/tnu-jst.11154

Email: phamhuongit@gmail.com

1. Giới thiệu

Phát hiện đối tượng là một trong những bài toán quan trọng của thị giác máy tính với ứng dụng đa dạng trong nhiều lĩnh vực như: công nghệ robot, xử lý ảnh y khoa, các hệ thống giám sát, hệ thống tương tác người-máy và giao thông thông minh. Trong công nghệ robot, phát hiện đối tượng giúp định vị và nhận dạng, cho phép robot tương tác chính xác với các vật thể trong thực tế. Trong xử lý ảnh y khoa, các ảnh chụp như X quang có thể được xử lý tự động để phát hiện các vùng bất thường, chẳng hạn như vùng chứa khối ung thư. Đối với các hệ thống giám sát, phát hiện đối tượng hỗ trợ việc nhận diện người, phương tiện, vật thể qua camera và dữ liệu này sẽ tiếp tục được xử lý để phục vụ các chức năng nâng cao. Trong các hệ thống tương tác người-máy, vị trí khuôn mặt hoặc cánh tay người sẽ được xác định bằng các thuật toán phát hiện đối tượng, sau đó được nhận dạng, phân tích để xác định chỉ thị cho máy. Trong giao thông thông minh, phát hiện đối tượng là yếu tố quan trọng trong xe tự hành, giúp nhận biết tự động các vật cản. Gần đây, các phương pháp phát hiện đối tượng đã có những tiến bộ lớn về độ chính xác và tốc độ xử lý. Nhiều nghiên cứu đã được đề xuất, từ việc sử dụng các đặc trưng thiết kế như Haar-like [1], HOG [2], DPM [3] trong các phương pháp truyền thống đến việc sử dụng các kỹ thuật hiện đại dựa trên mạng học sâu như SSD [4], Retinanet [5], R-CNN [6], Fast R-CNN [7], Faster R-CNN [8], Mask R-CNN [9], YOLO [10].

Trong lĩnh vực quân sự nói chung và ngành trinh sát nói riêng việc phát hiện các đối tượng được ngụy trang là một thách thức lớn đối với các hệ thống quân sự và an ninh. Ngụy trang là một kỹ thuật phổ biến được sử dụng để che giấu vị trí, giảm thiểu nguy cơ bị phát hiện và tăng cường khả năng tấn công hoặc ẩn nấp. Các đối tượng ngụy trang thường sử dụng màu sắc, hoa văn và cấu trúc để hòa lẫn vào môi trường xung quanh, làm cho chúng trở nên khó phát hiện đối với con người hoặc các hệ thống tự động phát hiện. Việc phát hiện đối tượng ngụy trang trong quân sự có ý nghĩa quan trọng trong việc bảo vệ lãnh thổ, đối phó với các hoạt động tình báo và đảm bảo an ninh quốc gia. Tính toàn vẹn của hệ thống phát hiện ngụy trang đóng vai trò quan trọng trong việc đảm bảo sự thành công của các hoạt động quân sự và chiến lược quốc gia. Trong nhiều năm qua, nghiên cứu và phát triển các phương pháp và công nghệ phát hiện ngụy trang đã đạt được sự tiến bộ đáng kể, nhưng vẫn còn nhiều thách thức cần được vượt qua. Các nhà nghiên cứu và các tổ chức quân sự trên toàn thế giới đang tiếp tục nỗ lực để phát triển các công nghệ phát hiện đối tượng ngụy trang hiệu quả và chính xác, giúp cải thiện khả năng quân sự và an ninh.

Một hướng tiếp cận hiệu quả là sử dụng kỹ thuật học sâu. Việc xây dựng mô hình tự động phát hiện ngụy trang trong quân sự là rất quan trọng và có nhiều ý nghĩa.

- Nâng cao khả năng phát hiện: Một mô hình tự động phát hiện ngụy trang có thể cung cấp khả năng phát hiện nhanh chóng và chính xác hơn so với phương pháp thủ công hoặc phụ thuộc vào con người. Điều này có thể giúp tăng cường khả năng phát hiện đối tượng ngụy trang và giảm thiểu rủi ro cho các hoạt động quân sự và an ninh.
- Giảm bớt áp lực, hỗ trợ cho lực lượng trinh sát: Việc áp dụng mô hình tự động phát hiện ngụy trang giúp giảm bớt áp lực và công sức cho lực lượng trinh sát và an ninh, đặc biệt tại những nơi đang có chiến sự, hoặc địa hình hiểm trở khó tiếp cận trực tiếp.
- Tăng cường an ninh quốc gia: Mô hình tự động phát hiện ngụy trang có thể đóng vai trò quan trọng trong việc bảo vệ lãnh thổ quốc gia, đối phó với các hoạt động tình báo và ngăn chặn các mối đe dọa tiềm ẩn từ các đối tượng ngụy trang.
- Tiết kiệm thời gian và chi phí: Sử dụng mô hình tự động phát hiện ngụy trang có thể giúp tiết kiệm thời gian và chi phí so với việc sử dụng các phương pháp thủ công, đồng thời tăng cường hiệu quả và chính xác trong việc phát hiện.
- Khả năng ứng dụng rộng rãi: Các mô hình tự động phát hiện ngụy trang có thể được áp dụng trong nhiều lĩnh vực khác nhau, bao gồm an ninh biên giới, giám sát an ninh công cộng và hỗ trợ cho các hoạt động quân sự và tình báo.

Tóm lại, việc xây dựng mô hình tự động phát hiện ngụy trang đóng vai trò quan trọng trong việc tăng cường an ninh quốc gia, giảm thiểu rủi ro và chi phí, cũng như nâng cao khả năng phát hiện và ứng dụng trong thực tế.

Hiện nay, có một số nghiên cứu đã được thực hiện về phát hiện ngụy trang trong quân sự, bao gồm cả các phương pháp sử dụng các mạng nơ-ron sâu và các kỹ thuật học máy khác. Hongbo Bi và các cộng sự [11] đã đưa ra một bài đánh giá toàn diện về các mô hình và bộ dữ liệu công khai trong các nghiên cứu về phát hiện đối tượng ngụy trang. Trong bài báo, các tác giả đã tóm tắt 39 mô hình COD từ năm 1998 đến 2021, phân loại chúng thành hai loại: 27 cấu trúc dựa trên đặc trưng thủ công và 12 cấu trúc dựa trên học sâu. Phân nhóm các cấu trúc thủ công thành sáu loại theo cơ chế phát hiện (kết cấu, màu sắc, chuyển động, cường độ, luồng quang học, và hợp nhất đa phương thức) và phân tích sâu các cấu trúc học sâu dựa trên động lực và hiệu suất phát hiện. Đồng thời cũng tổng hợp và mô tả chi tiết bốn bộ dữ liệu COD phổ biến, thảo luận về các hạn chế và giải pháp để cải thiện độ chính xác phát hiện, đồng thời đề xuất hướng nghiên cứu tương lai để phát triển COD.

Nghiên cứu của Maozhen Liu cùng cộng sự [12] giới thiệu một mạng nơ-ron mới được gọi là MHNet, được phát triển đặc biệt để phát hiện các đối tượng ngụy trang trong môi trường quân sự. Mục tiêu của mạng này là giúp cho việc phát hiện các đối tượng được ngụy trang quân sự trong các bối cảnh phức tạp trở nên hiệu quả hơn. Nhóm tác giả này đề cập đến việc xây dựng một bộ dữ liệu mới, được gọi là MH-COCO, để huấn luyện và đánh giá MHNet. Bộ dữ liệu này chứa các hình ảnh với đa dạng các đối tượng được ngụy trang quân sự trong các tình huống khác nhau, bao gồm cả các tình huống với nền phức tạp. MHNet được mô tả là một mạng nơ-ron tích chập (CNN) đặc biệt được thiết kế để xử lý các cảnh quân sự có đặc điểm khác biệt so với các cảnh quang học thông thường. Bằng cách sử dụng kiến trúc và kỹ thuật đặc biệt, MHNet có khả năng phát hiện các đối tượng ngụy trang một cách hiệu quả trong các tình huống phức tạp. Nghiên cứu đã cung cấp các kết quả thực nghiệm chi tiết để minh họa hiệu suất của MHNet so với các phương pháp phát hiện đối tượng khác trong các tình huống quân sự. Kết quả cho thấy rằng MHNet đạt được kết quả tốt và vượt trội so với các phương pháp hiện có trong các tình huống ngụy trang phức tạp.

Trong nghiên cứu này, chúng tôi đã sử dụng bộ dữ liệu công bố trong nghiên cứu của Maozhen Liu cùng cộng sự [12] và đề xuất áp dụng mô hình YOLOv9 [13] để tự động phát hiện đối tượng quân sự được ngụy trang từ hình ảnh. Kết quả của nghiên cứu này có thể được sử dụng trong trinh sát, giám sát an ninh,… Ngoài ra, bài báo cũng sẽ trình bày kết quả thực nghiệm và thảo luận về hiệu quả của phương pháp đề xuất, từ đó đề xuất hướng phát triển tiếp theo cho lĩnh vực này.

2. Dữ liệu và phương pháp

2.1. Bộ dữ liệu phát hiện đối tượng ngụy trang quân sự

Các bộ dữ liệu đóng vai trò then chốt trong các phương pháp phát hiện đối tượng dựa trên học máy. Tuy nhiên, cho tới hiện tại, có rất ít bộ dữ liệu về đối tượng ngụy trang công khai sẵn có, vì việc phát hiện các đối tượng ngụy trang thường được coi là một nhiệm vụ đặc biệt trong các thử nghiệm thị giác máy tính trong mấy năm gần đây. Về phát hiện đối tượng ngụy trang nói chung, chúng ta có thể kể đến 04 bộ dữ liệu có sẵn như sau:

- CHAMELEON [14]. Tập dữ liệu này bao gồm 76 hình ảnh tự nhiên, được thu thập từ Internet thông qua từ khóa "động vật ngụy trang". Mỗi hình ảnh của CHAMELEON đều có các nhãn được đánh dấu bằng tay tương ứng. Tập dữ liệu này chủ yếu tập trung vào các loài động vật ngụy trang trong các phông nền phức tạp mà hệ thống thị giác của con người gần như không thể phân biệt được mục tiêu và môi trường xung quanh. Do đó, tập dữ liệu này thích hợp để xác minh tính khả dụng của các mô hình.

- CAMO-COCO [15] được tạo thành từ hai tập dữ liệu con, một tập là hình ảnh đối tượng ngụy trang (CAMO), và một tập là hình ảnh đối tượng không ngụy trang (MS-COCO). Cả

CAMO và MS-COCO đều chứa 1.250 hình ảnh. Các hình ảnh trong tập dữ liệu CAMO bao gồm nhiều tình huống đầy thách thức như ngoại hình đối tượng, sự lộn xộn của nền, độ phức tạp của hình dạng, đối tượng nhỏ, che khuất đối tượng, nhiều đối tượng và sự mất tập trung. Đây là tập dữ liệu đầu tiên cho phân vùng ngụy trang, nhưng vẫn phù hợp cho phát hiện đối tượng ngụy trang vì vậy vẫn có thể sử dụng để xác minh tính khả dụng của các mô hình.

- Fan và cộng sự [16] đã đề xuất một bộ dữ liệu phát hiện đối tượng ngụy trang có số lượng hình ảnh lớn nhất - COD10K, bao gồm 10.000 hình ảnh được chia thành 10 lớp lớn và 78 lớp con (69 lớp ngụy trang và 9 lớp không ngụy trang). Ngoài ra, kích thước của các đối tượng trong COD10K được phân thành ba loại: lớn, trung bình, và nhỏ, giúp đánh giá chính xác hiệu quả của các mô hình theo từng khía cạnh. Tương tự như CAMO, COD10K cũng được gán nhãn từng đối tượng để tạo ra Ground Truth tương ứng cho mỗi hình ảnh. COD10K là tập dữ liệu phát hiện đối tượng ngụy trang lớn nhất cho đến nay, với các nhãn phong phú nhất, cho phép hiểu toàn diện về đối tượng ngụy trang và có thể được sử dụng để phát hiện, phân vùng đối tượng ngụy trang.

- Lyu và cộng sự [17] đã đề xuất một bộ dữ liệu thử nghiệm mới (NC4K) cho phát hiện đối tượng ngụy trang, bao gồm 4.121 hình ảnh thu thập từ Internet. Các cảnh ngụy trang của tập dữ liệu này có thể được chia thành hai loại chính: ngụy trang tự nhiên và ngụy trang nhân tạo. Hầu hết các hình ảnh trong bộ dữ liệu này thuộc loại ngụy trang tự nhiên, tức là các sinh vật ngụy trang bằng cách sử dụng các đặc điểm vật lý hoặc hóa học để hòa lẫn vào nền cảnh xung quanh.

Các bộ dữ liệu kể trên chủ yếu chỉ áp dụng cho phân đoạn ngữ nghĩa của cây cối và động vật. Cho đến nay, vẫn chưa có bộ dữ liệu nào dành riêng cho ngụy trang quân sự và đặc biệt là cho ngụy trang cấp cao. Để nâng cao hiệu quả của các phương pháp phát hiện đối tượng thì việc xây dựng các bộ dữ liệu chuyên biệt theo miền là cần thiết vì các đặc điểm của các đối tượng có sự khác biệt lớn giữa các miền khác nhau. Trong công trình nghiên cứu gần đây [18], cũng đã có bộ dữ liệu các đối tượng quân sự nhưng nó cũng chỉ dành cho các đối tượng quân sự chung (các đối tượng rõ ràng) chứ không phải cho ngụy trang cấp cao.

Do đó, trong nghiên cứu này, tôi sử dụng bộ dữ liệu chứa các đối tượng quân sự được ngụy trang cấp cao của nhóm tác giả Maozhen Liu và Xiaoguang Di [12] có sẵn trên trang web https://github.com/liumaozhen-lmz/Military-Camouflage-MHCD2022.git.

Hình 1. *Ảnh các đối tượng người, xe quân sự, tàu chiến ngụy trang đã được gán nhãn*

Để xây dựng bộ dữ liệu này, các tác giả đã thu thập hình ảnh từ trang web giống như trong [18]. Để có được mẫu ngụy trang thực tế, các tác giả đã thu thập các hình ảnh ngụy trang trong bối cảnh quân sự từ công cụ tìm kiếm web bằng cách thiết lập nhiều bộ từ khóa. Ngoài ra, các tác giả cũng thu thập một số đoạn video và bảo quản các khung hình liên tục với 27 FPS. Bộ dữ liệu bao gồm năm loại đối tượng: người, máy bay, xe quân sự, tàu chiến và xe tăng, với ngụy trang cấp cao liên quan đến nhiều kịch bản thực tế như rừng rậm, sa mạc, tuyết, thị trấn và đại dương. Sau đó, hợp nhất một bộ dữ liệu nhỏ gồm 1.000 ảnh ngụy trang của người đơn lẻ cho phân đoạn ngữ nghĩa [19] và gán nhãn lại bằng các hộp giới hạn. Bộ dữ liệu này rất phù hợp cho việc đánh giá các phương pháp phát hiện đối tượng ngụy trang tiên tiến cho quân sự.

Sau khi việc thu thập mẫu hoàn tất, việc làm sạch dữ liệu và lọc thủ công được thực hiện để loại bỏ các hình ảnh không đạt yêu cầu. Các hình ảnh không đạt tiêu chuẩn (mờ, độ phân giải thấp, không có đối tượng ngụy trang quân sự, trùng lặp cao,...) đã được loại bỏ trước khi gán nhãn. Bộ dữ liệu cuối cùng gồm 3.000 hình ảnh ngụy trang quân sự cấp cao đã được gán nhãn cẩn thận với các loại đối tượng bằng các hình chữ nhật bao quanh đối tượng (minh họa trong Hình 1).

2.2. Mô hình YOLOv9

Trong phần này, tôi sẽ trình bày chi tiết về kiến trúc của mô hình YOLOv9, một phiên bản tiên tiến nhất của dòng mô hình YOLO (You Only Look Once) [13]. Mô hình này được thiết kế để tối ưu hóa cả về tốc độ và độ chính xác trong phát hiện đối tượng, đặc biệt trong các tình huống phức tạp như phát hiện ngụy trang quân sự.

Kiến trúc YOLOv9, được thiết kế cho việc phát hiện đối tượng thời gian thực, có nhiều cải tiến so với các mô hình YOLO trước đây. Phần này chi tiết các thành phần chính và các yếu tố làm cho YOLOv9 trở thành công cụ mạnh mẽ để phát hiện các đối tượng quân sự ngụy trang. Hai kỹ thuật chính được đề xuất trong YOLOv9 là thông tin gradient lập trình (PGI - Programmable Gradient Information) và kiến trúc GELAN (Generalized Efficient Layer Aggregation Network). PGI có thể cung cấp thông tin đầu vào đầy đủ cho nhiệm vụ mục tiêu để tính toán hàm mục tiêu, từ đó thu được thông tin gradient đáng tin cậy để cập nhật trọng số mạng. Ngoài ra, một kiến trúc mạng nhẹ mới có tên gọi là GELAN, dựa trên quy hoạch đường dẫn gradient được giới thiệu. Kết quả thử nghiệm cho thấy GELAN chỉ sử dụng các toán tử tích chập thông thường để đạt được hiệu suất sử dụng tham số tốt hơn các phương pháp tiên tiến được phát triển dựa trên tích chập theo chiều sâu. PGI có thể được sử dụng cho nhiều loại mô hình từ nhẹ đến lớn. Nó có thể được sử dụng để thu thập thông tin đầy đủ, do đó các mô hình huấn luyện từ đầu có thể đạt được kết quả tốt hơn so với các mô hình tiên tiến được huấn luyện trước bằng các tập dữ liệu lớn.

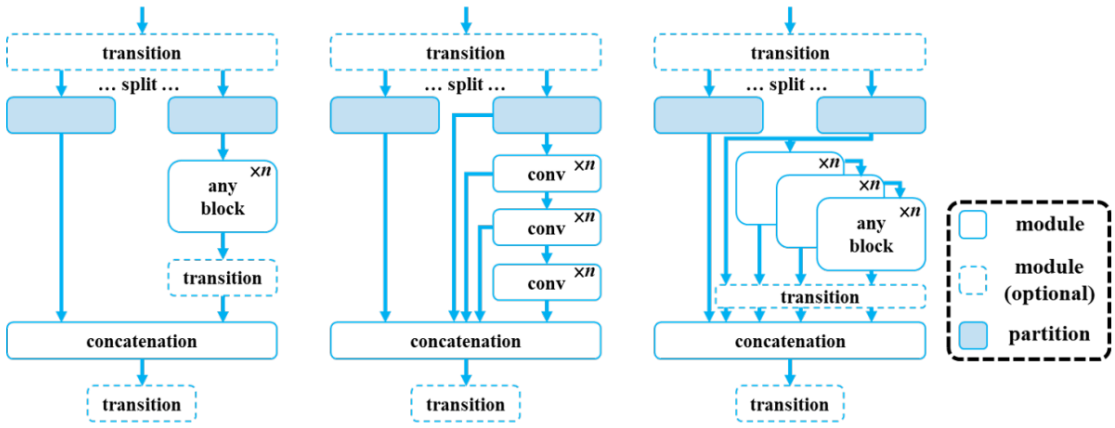
Cơ chế PGI trong YOLOv9 giúp nâng cao chất lượng gradient, giải quyết các vấn đề như biến mất và bùng nổ gradient. Nó đảm bảo luồng gradient đáng tin cậy và toàn diện, cải thiện hiệu suất tổng thể của mô hình. Các thành phần chính của PGI gồm có:

- Nhánh phụ trợ có thể đảo ngược: Duy trì thông tin đầy đủ thông qua kiến trúc có thể đảo ngược, đảm bảo gradient chảy ngược mà không bị suy giảm.
- Thông tin phụ trợ đa mức: Phân phối thông tin phụ trợ qua các mức độ khác nhau để giảm thiểu sự tích lũy lỗi trong giám sát sâu.
- Lập trình thông tin gradient: Tăng cường và lập trình luồng gradient để tối ưu hóa hiệu quả huấn luyện và đảm bảo gradient đáng tin cậy cho nhánh chính.

Trong YOLOv9, quy trình huấn luyện sử dụng PGI để nâng cao chất lượng gradient, sử dụng nhánh phụ trợ có thể đảo ngược và thông tin phụ trợ đa cấp. Các hàm mất mát bổ sung cho PGI, tạo ra gradient mạnh mẽ và đáng tin cậy.

Kiến trúc mạng GELAN kết hợp các ưu điểm của CSPNet (Cross Stage Partial Network) và ELAN (Efficient Layer Aggregation Network) (minh họa trong Hình 2). Kiến trúc GELAN được thiết kế để cải thiện việc sử dụng tham số và duy trì sự cân bằng giữa độ phức tạp của mô hình và hiệu quả tính toán. CSPNet tăng cường luồng gradient qua mạng bằng cách tích hợp các kết nối giai đoạn chéo. Điều này giảm chi phí tính toán và cải thiện khả năng học tập của mạng bằng

cách phân chia bản đồ đặc trưng của lớp cơ sở thành hai phần và sau đó hợp nhất chúng qua một hệ thống phân cấp giai đoạn chéo. Các khối ELAN được sử dụng để tổng hợp các đặc trưng một cách hiệu quả. Những khối này xếp chồng nhiều lớp tích chập, giúp thu thập một loạt các đặc trưng ở các mức độ trừu tượng khác nhau. Trong YOLOv9, khối ELAN truyền thống được mở rộng để hỗ trợ bất kỳ khối tính toán nào, làm cho kiến trúc linh hoạt hơn.



(a) CSPNet *(b) ELAN* *(c) GELAN*

Hình 2. *Kiến trúc của GELAN: (a) CSPNet, (b) ELAN và (c) GELAN được đề xuất cho YOLOv9* [13]

Bảng 1. *Cấu hình mạng nơ-ron của YOLOv9*

STT	Module	Filters	Size	Stride
1	Conv	64	3	2
2	Conv	128	3	2
3	GELAN	256, 128, 64	–	1
4	DOWN	256	3	2
5	CSP-ELAN	512, 256, 128	–	1
6	DOWN	512	3	2
7	CSP-ELAN	512, 512, 256	–	1
8	DOWN	512	3	2
9	CSP-ELAN	512, 512, 256	–	1
10	SPP-ELAN	512, 256, 256	–	1
11	Up	512	–	2
12	Concat	1024	–	1
13	CSP-ELAN	512, 512, 256	–	1
14	Up	512	–	2
15	Concat	1024	–	1
16	CSP-ELAN	256, 256, 128	–	1
17	DOWN	256	3	2
18	Concat	768	–	1
19	CSP-ELAN	512, 512, 256	–	1
20	DOWN	512	3	2
21	Concat	1024	–	1
22	CSP-ELAN	512, 512, 256	–	1

Kiến trúc tổng thể của YOLOv9 tương tự như kiến trúc của YOLOv7, được cải tiến bằng cách dùng các khối GELAN thay thế cho các khối ELAN của YOLOv7 và áp dụng cơ chế PGI trong quá trình huấn luyện. Thông số cụ thể của các thành phần của mạng nơ-ron được mô tả ở Bảng 1.

Mô hình YOLOv9 là một bước tiến đáng kể trong phát hiện đối tượng thời gian thực, đặc biệt cho việc phát hiện các đối tượng quân sự ngụy trang. Việc sử dụng sáng tạo kiến trúc GELAN,

cơ chế PGI, kết hợp với các chiến lược huấn luyện và tối ưu hóa hiệu quả, khiến YOLOv9 trở thành một mô hình hiệu quả và linh hoạt cho các nhiệm vụ phát hiện đối tượng đầy thách thức.

3. Thử nghiệm, đánh giá

3.1. Phương pháp đánh giá mô hình

Để đánh giá toàn diện kết quả của các mô hình phát hiện đối tượng, chúng tôi đã sử dụng chỉ số mAP50. Độ chính xác trung bình mAP (mean Average Precision) là chỉ số chính được sử dụng để đánh giá các mô hình phát hiện đối tượng. Nó được tính bằng trung bình của Độ chính xác AP (Average Precision) trên tất cả các lớp. AP đo lường sự đánh đổi giữa chỉ số Precision và chỉ số Recall cho mỗi lớp bằng cách tích hợp đường cong precision-recall. Cụ thể, mAP tại ngưỡng 50% của tỉ lệ giữa diện tích phần giao chia cho diện tích phần hợp giữa kết quả dự đoán và nhãn (IoU - Intersection over Union), ký hiệu là mAP50, được sử dụng trong nghiên cứu này. Nó đại diện cho độ chính xác trung bình khi các bounding boxes dự đoán có ít nhất 50% chồng lấn với các bounding boxes thực tế.

$$\text{mAP50} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_{50,i} \quad (1)$$

Trong đó N là số lượng lớp đối tượng, và $\text{AP}_{50,\text{i}}$ là độ chính xác trung bình tại IoU 50% cho lớp i.

3.2. Thiết lập thử nghiệm

Tập dữ liệu được sử dụng trong nghiên cứu này bao gồm 3.000 hình ảnh chứa các đối tượng quân sự ngụy trang trong nhiều bối cảnh khác nhau. Các lớp đối tượng bao gồm Người, Máy bay, Xe quân sự, Tàu chiến và Xe tăng. Dữ liệu được gán nhãn (đánh dấu đối tượng bằng các hình chữ nhật bao quanh đối tượng) và được chia thành 3 tập dữ liệu độc lập cụ thể như sau: tập huấn luyện (training set): 1.920 ảnh; tập đánh giá (validation set): 480 ảnh và tập kiểm tra (test set): 600 ảnh.

Trong nghiên cứu này đã triển khai và huấn luyện các mô hình phát hiện đối tượng khác nhau, bao gồm SSD [20], Faster R-CNN [21], DETR [22], FCOS [23], Retinanet [24], Cascade R-CNN [25], Sparse R-CNN [26], Decoupled R-CNN [27], ERFNet [28], PPYOLOE-M [29], AdaMixer [30], CrossDet++ [31], MHNet [12], YOLOv7 [32] và YOLOv9. Sau khi huấn luyện, các mô hình được kiểm tra trên tập kiểm tra để đánh giá hiệu suất. Kết quả mAP50 của các mô hình được so sánh để xác định mô hình hiệu quả nhất trong việc phát hiện các đối tượng quân sự ngụy trang.

Quá trình huấn luyện và kiểm tra các mô hình được thực hiện trên một hệ thống máy tính với cấu hình phần cứng như sau: GPU NVIDIA RTX 1080 với bộ nhớ 12GB, RAM 96 GB, CPU 12 cores.

Trong thử nghiệm, khi huấn luyện mô hình YOLOv9 đã thiết lập các tham số nhằm điều chỉnh quá trình học và tối ưu hóa mô hình: sử dụng trọng số mô hình đã được huấn luyện trước yolov9-e-converted.pt; số epochs: 150; batch_size: 8; images size: 640; optimizer: SGD (bộ tối ưu hóa được sử dụng là SGD - Stochastic Gradient Descent); lr0: 0.01 (learning rate ban đầu); lrf: 0.01 (tỷ lệ giảm learning rate). Trong thử nghiệm này cũng đã áp dụng các kỹ thuật tăng cường dữ liệu để cải thiện hiệu suất tổng thể như: điều chỉnh Hue (sắc độ màu), Saturation (độ bão hòa màu), Value (độ sáng) của hình ảnh, phóng to, thu nhỏ, áp dụng kỹ thuật mosaic (ghép nhiều ảnh lại để tạo ra một ảnh mới), sử dụng kỹ thuật mixup (trộn hai ảnh ngẫu nhiên).

3.3. Kết quả thực nghiệm

Các mô hình phát hiện đối tượng khác nhau được đánh giá trên tập dữ liệu của nhóm tác giả Maozhen Liu và Xiaoguang Di [12] bao gồm các hình ảnh của các đối tượng quân sự ngụy trang trong nhiều bối cảnh khác nhau. Các mô hình được thử nghiệm bao gồm SSD, Faster R-CNN, DETR, FCOS, Retinanet, Cascade R-CNN, Sparse R-CNN, Decoupled R-CNN, ERFNet, PPYOLOE-M, AdaMixer, CrossDet++, MHNet [12] và YOLOv9. Bảng 2 tóm tắt kết quả đánh giá phát hiện trên các loại đối tượng khác nhau: Người, Máy bay, Xe quân sự, Tàu chiến, và Xe tăng, cùng với chỉ số mAP50.

Bảng 2. *Kết quả phát hiện đối tượng ngụy trang trên tập dữ liệu thử nghiệm*

Mô hình	Người	Máy bay	Xe quân sự	Tàu chiến	Xe tăng	mAP50
SSD [20]	31,85	68,98	38,59	39,94	60,65	48,00
Faster R-CNN [21]	66,24	63,91	33,61	40,17	63,86	53,56
DETR [22]	65,72	67,04	39,73	45,53	62,29	56,58
FCOS [23]	63,15	65,21	32,78	42,32	60,19	52,73
Retinanet [24]	64,40	63,68	34,73	41,56	61,16	53,12
Cascade R-CNN [25]	66,82	64,31	31,08	43,57	62,24	53,60
Sparse R-CNN [26]	62,39	62,57	30,56	42,34	61,78	51,93
Decoupled R-CNN [27]	65,74	67,25	38,91	46,22	64,29	55,82
ERFNet [28]	64,32	66,71	39,17	45,51	63,58	55,78
PPYOLOE-M [29]	65,26	63,81	38,65	46,49	63,41	55,98
AdaMixer [30]	54,55	69,79	39,67	47,80	64,07	56,83
CrossDet++ [31]	66,98	64,79	39,71	46,59	64,10	56,42
MHNet [12]	67,37	65,22	40,03	46,87	64,31	56,79
YOLOv7 [32]	81,1	**72,6**	27,4	58,8	63,4	60,7
YOLOv9 [13]	**88,6**	50,4	**41,4**	**73,7**	**76,1**	**66,1**

Kết quả cho thấy mô hình YOLOv9 vượt trội trong việc phát hiện các đối tượng quân sự ngụy trang, đạt được mAP50 ấn tượng là 66,1%, cao hơn đáng kể so với các mô hình khác được thử nghiệm. Đặc biệt, YOLOv9 xuất sắc trong việc phát hiện các lớp "Người" và "Xe tăng", với tỷ lệ phát hiện lần lượt là 88,6% và 76,1%. Điều này cho thấy khả năng mạnh mẽ của mô hình trong việc nhận diện các đối tượng khó phát hiện, vốn thường gặp khó khăn do sự phức tạp của ngụy trang.

Mặc dù YOLOv7 cho thấy hiệu suất mạnh mẽ với mAP50 cao là 60,7%, nhưng nó gặp khó khăn trong việc phát hiện Xe quân sự, chỉ đạt 27,4%. Điều này cho thấy điểm yếu của YOLOv7 trong việc phát hiện một số loại phương tiện quân sự, có thể cần tối ưu hóa thêm hoặc tăng cường dữ liệu.

MHNet đạt được hiệu suất khá ấn tượng với mAP50 là 56,79%, đứng thứ ba trong bảng xếp hạng tổng thể. Điều này cho thấy rằng MHNet là một mô hình phát hiện đối tượng mạnh mẽ, đặc biệt là trong việc phát hiện các đối tượng ngụy trang quân sự.

Các mô hình khác, như Faster R-CNN và DETR, cũng cho thấy hiệu suất hợp lý với mAP50 lần lượt là 53,56% và 56,58%. Những mô hình này thể hiện hiệu suất cân bằng trên các danh mục khác nhau nhưng bị YOLOv9 vượt trội trong bối cảnh cụ thể của việc phát hiện đối tượng ngụy trang.

Mô hình YOLOv9 đã chứng minh hiệu suất tốt trong việc phát hiện các đối tượng quân sự ngụy trang, vượt trội so với các mô hình học sâu khác về độ chính xác. Nghiên cứu này không chỉ xác nhận tính khả thi của việc áp dụng học sâu vào nhiệm vụ đầy thách thức này mà còn mở ra hướng đi mới cho các nghiên cứu và ứng dụng trong lĩnh vực này, nâng cao tiềm năng về an ninh và hiệu quả hoạt động trong các ứng dụng quân sự.

4. Kết luận

Bài báo này đã nghiên cứu việc ứng dụng các kỹ thuật học sâu, cụ thể là mô hình YOLOv9, cho việc phát hiện các đối tượng quân sự ngụy trang. Thách thức trong việc nhận diện các đối tượng ngụy trang, vốn được thiết kế để hòa lẫn vào môi trường xung quanh, đã được giải quyết bằng cách tận dụng các khả năng tiên tiến của các mô hình phát hiện đối tượng hiện đại.

Mô hình YOLOv9 đã thể hiện hiệu suất vượt trội so với một số mô hình phát hiện đối tượng tiên tiến khác. YOLOv9 đạt được chỉ số mAP50 cao nhất trong tất cả các mô hình được thử nghiệm, cao hơn mô hình đứng thứ hai 9,27%. Kết quả này cho thấy độ chính xác vượt trội của nó trong việc phát hiện các đối tượng ngụy trang.

Dựa trên những kết quả này, có thể mở ra một số hướng nghiên cứu trong tương lai để tiếp tục nâng cao tính chính xác, ổn định của mô hình. Hướng thứ nhất là mở rộng tập dữ liệu để bao gồm nhiều kịch bản ngụy trang đa dạng và thách thức hơn, có thể cải thiện thêm tính ổn định và khả

năng tổng quát của mô hình. Có thể kết hợp YOLOv9 với các dữ liệu cảm biến khác và hệ thống máy học có thể nâng cao khả năng của nó trong các ứng dụng thực tế phức tạp. Mặc dù YOLOv9 đã thể hiện hiệu suất tốt, việc khám phá các kỹ thuật học sâu mới có thể mang lại những cải tiến thêm về độ chính xác và hiệu quả.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.

[2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893, 2005.

[3] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of 14th European Conference on Computer Vision–ECCV 2016*, Amsterdam, The Netherlands, Part I, 2016, pp. 21-37.

[5] T. Lin, "Focal loss for dense object detection," *arXiv preprint arXiv*:1708.02002, 2017.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580-587.

[7] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision*, 2015, pp. 1440-1448.

[8] S. Ren, "Faster RCNN: towards real-time object detection with region proposal networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2015, pp. 91-99.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2961-2969.

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779-788.

[11] H. Bi, C. Zhang, K. Wang, J. Tong, and F. Zheng, "Rethinking camouflaged object detection: Models and datasets," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5708-5724, 2021.

[12] M. Liu, and X. Di, "Extraordinary MHNet: Military high-level camouflage object detection network and dataset," *Neurocomputing*, vol. 549, p. 126466, 2023.

[13] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv*:2402.13616, 2014.

[14] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, "Animal camouflage analysis: Chameleon database," 2017. [Online]. Available: https://www.polsl.pl/rau6/chameleon-database-animal-camouflage-analysis/. [Accessed Nov. 8, 2024].

[15] T. N. Le, T. V. Nguyen, Z. Nie, M. T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Computer Vision and Image Understanding*, vol. 184, pp. 45-56, 2019.

[16] D. P. Fan, G.P. Ji, M. M. Cheng, and L. Shao, "Concealed object detection," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6024-6042, 2021.

[17] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11591-11601.

[18] X. Yi, J. Wu, B. Ma, Y. Ou, and L. Liu, "MOD: benchmark for military object detection," *arXiv preprint arXiv*:2104.13763, 2021.

[19] Y. Zheng, X. Zhang, F. Wang, T. Cao, M. Sun, and X. Wang, "Detection of people with camouflage pattern via dense deconvolution network," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 29-33, 2018.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "Ssd: Single shot multibox detector," In *Proceedings of 14th European Conference on Computer Vision–ECCV 2016*, Amsterdam, The Netherlands, Part I 14, 2016, pp. 21-37.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137-1149, 2016.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," In *European conference on computer vision*, 2020, pp. 213-229.

[23] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," *arXiv preprint arXiv:1904.01355*, 2019.

[24] T. Y. Ross and G. K. H. P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2980-2988.

[25] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154-6162.

[26] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang, and P. Luo, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14454-14463.

[27] D. Wang, K. Shang, H. Wu and C. Wang, "Decoupled R-CNN: Sensitivity-specific detector for higher accurate localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6324-6336, 2022.

[28] Q. Wang, S. Zhang, Y. Qian, G. Zhang, and H. Wang, "Enhancing representation learning by exploiting effective receptive fields for object detection," *Neurocomputing*, vol. 481, pp. 22-32, 2022.

[29] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, and B. Lai, "PP-YOLOE: An evolved version of YOLO," *arXiv preprint arXiv:2203.16250*, 2022.

[30] Z. Gao, L. Wang, B. Han, and S. Guo, "Adamixer: A fast-converging query-based object detector," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5364-5373.

[31] H. Qiu, H. Li, Q. Wu, J. Cui, Z. Song, L. Wang, and M. Zhang, "CrossDet++: Growing crossline representation for object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1093-1108, 2022.

[32] C.Y. Wang, A. Bochkovskiy, and H.Y.M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7464-7475.