p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

ĐÁNH GIÁ HIỆU QUẢ CÁC MÔ HÌNH PHÁT HIỆN ĐỐI TƯỢNG DỰA TRÊN CƠ CHẾ HỢP NHẤT ĐẶC TRƯNG TRONG BỐI CẢNH ẢNH UAV

Nguyễn Dũng[1], Nguyễn Ngọc Thủy[1], Bùi Lương Vũ Ngọc[2]

TÓM TẮT

Phát hiện đối tượng từ góc nhìn UAV ngày càng được quan tâm nhờ các ứng dụng như giám sát giao thông, nông nghiệp thông minh và quan sát môi trường. Tuy nhiên, ảnh UAV thường chứa các đối tượng nhỏ, mật độ cao, bị che khuất và có nền phức tạp, gây nhiều thách thức. Bài báo này khảo sát và đánh giá thực nghiệm các mô hình phát hiện đối tượng hiện đại dựa trên CNN và Transformer trong kịch bản UAV, trên các bộ dữ liệu VisDrone2019, TinyPerson và HIT-UAV. Kết quả cho thấy sự đánh đổi rõ rệt giữa độ chính xác và chi phí tính toán, đồng thời các phương pháp chú ý thích ứng và tháp đặc trưng đa mức cho thấy tiềm năng cân bằng hiệu năng và hiệu quả cho triển khai UAV.

Từ khóa: *UAV, phát hiện đối tượng, học sâu, cơ chế chú ý, tháp đặc trưng đa mức.*

DOI: *https://doi.org/10.70117/hdujs.84.2.2026.1146*

1. ĐẶT VẤN ĐỀ

Sự phát triển nhanh chóng của thiết bị bay không người lái (Unmanned Aerial Vehicle - UAV) đã mở ra nhiều hướng ứng dụng mới trong lĩnh vực thị giác máy tính, trong đó phát hiện đối tượng (object detection) là một nhiệm vụ cốt lõi. So với các hệ thống camera quan sát mặt đất truyền thống, ảnh thu thập từ UAV có những đặc điểm riêng biệt như góc nhìn từ trên cao (bird's-eye view), độ phân giải không gian thay đổi đáng kể theo độ cao bay và thường chứa số lượng lớn các đối tượng có kích thước nhỏ. Theo các thống kê được báo cáo trên bộ dữ liệu VisDrone2019 [1], phần lớn các đối tượng trong ảnh UAV thuộc nhóm đối tượng nhỏ theo định nghĩa của COCO (diện tích <32x32 = 1024 pixel^2), với tỷ lệ ước tính trên 65%. Con số này cao hơn đáng kể so với các bộ dữ liệu phát hiện đối tượng phổ biến như COCO [2], nơi các đối tượng nhỏ chỉ chiếm khoảng 41% tổng số đối tượng. Đáng chú ý, VisDrone còn chứa một tỷ lệ lớn các đối tượng cực nhỏ (tiny objects, kích thước <16x16 pixel), chiếm xấp xỉ 35-40% trong khi tỷ lệ này trong COCO chỉ vào khoảng 12-15%. Sự chênh lệch rõ rệt về phân bố kích thước đối tượng này khiến các mô hình phát hiện đối tượng truyền thống gặp nhiều khó khăn khi được áp dụng trực tiếp cho các kịch bản UAV. Ba thách thức chính trong bài toán phát hiện đối tượng từ ảnh UAV có thể được tổng hợp như sau: (i) Đối tượng kích thước nhỏ và cực nhỏ: kích thước đối tượng rất hạn chế (thường <32×32 pixel) đặc trưng phân biệt do thiếu thông tin chi tiết hình học và ngữ cảnh; (ii) Mật độ đối tượng cao: số lượng lớn đối tượng tập trung trong một vùng không gian hẹp gây ra hiện tượng che khuất, chồng lấp và

[1] *Trường Đại học Khoa học, Đại học Huế; Email: nguyendung@hueuni.edu.vn*

[2] *Phân hiệu Trường Đại học Y Hà Nội tại tỉnh Thanh Hoá*

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

nhầm lẫn không gian giữa các đối tượng; (iii) Độ phức tạp của nền và độ tương phản thấp: nền ảnh đa dạng và phức tạp, cùng với độ tương phản thấp giữa đối tượng và môi trường xung quanh, làm gia tăng khó khăn trong việc phân tách giữa đối tượng và cảnh nền.

Trước bối cảnh đó, bài báo này tập trung vào việc đánh giá một cách có hệ thống các kiến trúc phát hiện đối tượng hiện đại trong kịch bản UAV. Các đóng góp chính của nghiên cứu bao gồm: (1) Khảo sát toàn diện các mô hình dựa trên CNN (Convolutional Neural Networks) và Transformer cho bài toán phát hiện đối tượng từ ảnh UAV, bao gồm các kiến trúc hai giai đoạn; các mô hình một giai đoạn; cũng như các mô hình dựa trên Transformer; (2) Phân tích chuyên sâu các chiến lược hợp nhất đặc trưng đa tỷ lệ như FPN, PANet, BiFPN và vai trò then chốt của chúng trong việc cải thiện hiệu suất phát hiện đối tượng nhỏ; (3) So sánh định lượng toàn diện giữa độ chính xác, độ phức tạp tính toán thông qua các chỉ số FLOPs (Floating-Point Operations Per Second) và số lượng tham số, cũng như khả năng xử lý thời gian thực được đánh giá bằng FPS (frames per second).

2. NGHIÊN CỨU LIÊN QUAN

2.1. Mô hình một giai đoạn

Các mô hình phát hiện một giai đoạn, tiêu biểu là dòng YOLO (You Only Look Once) [3], được phát triển với mục tiêu cân bằng giữa tốc độ suy luận và độ chính xác, và thường được đánh giá chủ yếu trên COCO. Sự tiến hóa của dòng YOLO phản ánh các nỗ lực nhằm cải thiện hiệu suất của mô hình một giai đoạn, đặc biệt trong việc xử lý các đối tượng có kích thước nhỏ, vốn là một thách thức chung trên cả COCO và dữ liệu UAV. YOLOv8 áp dụng thiết kế anchor-free [4] kết hợp với khối C2f (Cross-Stage Partial Feature Fusion), qua đó giảm sự phụ thuộc vào các anchor được thiết kế thủ công. Cách tiếp cận này giúp mô hình linh hoạt hơn trước sự thay đổi về kích thước và tỷ lệ đối tượng, đặc điểm thường gặp trong các kịch bản UAV do sự thay đổi độ cao và góc nhìn. YOLOv10 giới thiệu cơ chế PGI (Programmable Gradient Information) [5] nhằm cải thiện quá trình lan truyền gradient trong huấn luyện. Cách tiếp cận này được đề xuất để tăng cường khả năng học đặc trưng cho các đối tượng nhỏ, vốn chỉ chiếm một phần nhỏ diện tích ảnh và thường nhận tín hiệu huấn luyện yếu hơn. Phiên bản gần đây hơn, YOLOv11 [6], tích hợp các khối C3k2 và cơ chế chú ý C2PSA nhằm tăng cường khả năng mô hình hóa thông tin không gian và kênh trong bối cảnh nền phức tạp. Theo các báo cáo thực nghiệm trên COCO, các thay đổi kiến trúc này góp phần cải thiện hiệu suất đối với các đối tượng nhỏ, thường được phản ánh qua chỉ số mAPs dành cho nhóm này. Khi áp dụng cho các bộ dữ liệu UAV như VisDrone2019, nơi tỷ lệ đối tượng nhỏ cao hơn so với COCO, các thiết kế tương tự cũng được ghi nhận là mang lại cải thiện nhất định về hiệu suất, cho thấy khả năng chuyển giao của các cải tiến kiến trúc từ COCO sang bối cảnh UAV.

2.2. Mô hình hai giai đoạn

Bên cạnh các mô hình một giai đoạn, các mô hình hai giai đoạn như Faster R-CNN [7] và Cascade R-CNN [8] thường được sử dụng như chuẩn tham chiếu về độ chính xác trên COCO, đặc biệt trong các kịch bản yêu cầu định vị chính xác các đối tượng nhỏ hoặc bị che khuất. Các kiến trúc này hoạt động bằng cách sinh các vùng đề xuất ở giai đoạn đầu, sau đó

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

thực hiện phân loại và hồi quy bounding box ở giai đoạn sau. Cascade R-CNN [8] mở rộng ý tưởng này bằng cách áp dụng nhiều giai đoạn tinh chỉnh liên tiếp với các ngưỡng IoU (Intersection over Union) tăng dần, nhằm cải thiện độ chính xác ở các mức IoU cao. Trong nghiên cứu UAV, các mô hình hai giai đoạn thường được sử dụng như một mốc tham chiếu để đánh giá hiệu suất của các kiến trúc nhẹ hơn. Khi một mô hình một giai đoạn hoặc dựa trên Transformer đạt hiệu suất tiệm cận các mô hình hai giai đoạn trên cùng một bộ dữ liệu, điều đó cho thấy khả năng trích xuất đặc trưng và dự đoán của kiến trúc đó ở mức cạnh tranh. Tuy nhiên, chi phí tính toán của các mô hình hai giai đoạn thường cao, với số lượng phép toán lớn khi đánh giá trên COCO. Điều này gây hạn chế trong các kịch bản yêu cầu xử lý thời gian thực hoặc triển khai trên nền tảng UAV với tài nguyên tính toán, bộ nhớ và năng lượng hạn chế. Do đó, trong thực tế, các mô hình hai giai đoạn thường phù hợp hơn với các bài toán xử lý ngoại tuyến hoặc các hệ thống có đủ tài nguyên, trong khi các ứng dụng UAV thời gian thực ưu tiên các kiến trúc gọn nhẹ hơn.

2.3. Mô hình dựa trên Transformer

DETR (Detection TRansformer) [9] đưa ra một cách tiếp cận khác cho bài toán phát hiện đối tượng bằng cách mô hình hóa việc phát hiện như một bài toán dự đoán tập hợp, qua đó loại bỏ bước hậu xử lý NMS (Non-Maximum Suppresion) thường thấy trong các kiến trúc truyền thống. Trên COCO, DETR cho thấy khả năng xử lý các cảnh phức tạp với nhiều đối tượng thông qua cơ chế tự chú ý toàn cục, cho phép mô hình học các mối quan hệ không gian giữa các đối tượng. Tuy nhiên, DETR nguyên bản gặp hạn chế trong việc phát hiện các đối tượng nhỏ, chủ yếu do khả năng khai thác thông tin đa tỷ lệ còn hạn chế và chi phí tính toán cao của cơ chế chú ý toàn cục. Các biến thể như Deformable DETR [10] được đề xuất nhằm giải quyết vấn đề này bằng cách sử dụng cơ chế chú ý biến dạng đa tỷ lệ, trong đó mô hình chỉ tập trung vào một tập con các vị trí quan trọng trên bản đồ đặc trưng. Cách tiếp cận này giúp cải thiện hiệu suất đối với đối tượng nhỏ trên COCO đồng thời giảm chi phí tính toán, và được xem là phù hợp hơn với các kịch bản có phân bố đối tượng không đồng đều như dữ liệu UAV. RT-DETR (Real time DETR) [11] tiếp tục hướng nghiên cứu này với mục tiêu tiệm cận yêu cầu thời gian thực trên COCO. Mô hình kết hợp mạng CNN để trích xuất đặc trưng cục bộ hiệu quả với bộ giải mã của Transformer có độ phức tạp thấp hơn, đồng thời áp dụng chiến lược lựa chọn đối tượng nhằm giảm số lượng phép toán cần thiết. Cách thiết kế lai này cho phép khai thác ưu điểm của cả CNN và Transformer trong một kiến trúc thống nhất. Việc RT-DETR đạt được sự cân bằng giữa độ chính xác và chi phí tính toán trên COCO thường được xem là một chỉ báo về khả năng mở rộng sang các kịch bản triển khai thực tế, bao gồm UAV, nơi các ràng buộc về tài nguyên còn khắt khe hơn. Do đó, RT-DETR không chỉ đại diện cho một mô hình cụ thể, mà còn phản ánh một hướng tiếp cận trong thiết kế các kiến trúc Transformer có khả năng triển khai trong điều kiện thời gian thực.

2.4. Vai trò của COCO trong nghiên cứu phát hiện đối tượng UAV

Việc sử dụng COCO làm trục phân tích trong nghiên cứu phát hiện đối tượng UAV mang lại một số lợi ích chính. Thứ nhất, COCO là môi trường nơi nhiều cải tiến kiến trúc quan trọng được đề xuất và đánh giá, cho phép quan sát sự tiến hóa của mô hình qua các thế hệ kiến trúc khác nhau. Thứ hai, COCO cung cấp một giao thức đánh giá thống nhất, hỗ trợ

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

so sánh tương đối công bằng giữa các mô hình được phát triển bởi các nhóm nghiên cứu khác nhau. Thứ ba, việc phân tích hiệu suất trên COCO giúp nhận diện các hướng thiết kế có tính tổng quát cao, trước khi tiến hành điều chỉnh và tối ưu hóa cho các bộ dữ liệu UAV cụ thể. Theo cách tiếp cận này, nghiên cứu phát hiện đối tượng UAV có thể được xem là sự kế thừa và mở rộng có chọn lọc từ các nguyên lý thiết kế đã được kiểm chứng trên COCO. Các kỹ thuật như hợp nhất đặc trưng đa tỷ lệ, cơ chế chú ý, hay thiết kế anchor-free đều cho thấy khả năng ứng dụng trong bối cảnh UAV. Tuy nhiên, việc chuyển giao này đòi hỏi sự điều chỉnh phù hợp để đáp ứng các đặc thù của dữ liệu UAV, bao gồm góc nhìn từ trên cao, phân bố đối tượng không đồng đều và các ràng buộc nghiêm ngặt về tài nguyên tính toán.

3. PHÂN TÍCH CÁC CƠ CHẾ HỢP NHẤT ĐẶC TRƯNG

Feature Pyramid Network (FPN) [12] là một kiến trúc nền tảng được đề xuất nhằm giải quyết bài toán phát hiện đối tượng đa tỷ lệ trong thị giác máy tính. FPN xây dựng một kim tự tháp đặc trưng bằng cách kết hợp các đặc trưng tầng sâu (giàu ngữ nghĩa nhưng độ phân giải thấp) với các đặc trưng tầng nông (giàu chi tiết nhưng yếu ngữ nghĩa) thông qua cơ chế top-down pathway và lateral connections. Nhờ đó, mỗi mức trong kim tự tháp đều sở hữu đặc trưng có tính ngữ nghĩa mạnh, đồng thời vẫn duy trì thông tin không gian cần thiết cho việc phát hiện các đối tượng ở nhiều kích thước khác nhau. FPN đã trở thành thành phần "neck" tiêu chuẩn trong nhiều hệ thống phát hiện đối tượng hiện đại như Faster R-CNN [7], Mask R-CNN [13] và RetinaNet [14]. Tuy nhiên, trong các kịch bản phát hiện đối tượng nhỏ từ ảnh UAV, FPN truyền thống vẫn bộc lộ nhiều hạn chế. Thứ nhất, các đối tượng nhỏ thường chỉ chiếm rất ít pixel, khiến đặc trưng tầng sâu bị suy giảm nghiêm trọng. Thứ hai, sự chênh lệch ngữ nghĩa giữa các tầng nông và tầng sâu gây khó khăn cho quá trình hợp nhất đặc trưng. Thứ ba, bối cảnh nền phức tạp và độ tương phản thấp trong ảnh UAV làm giảm khả năng phân biệt giữa mục tiêu và nền. Do đó, nhiều nghiên cứu gần đây đã tập trung cải tiến FPN theo các hướng như tăng cường luồng thông tin hai chiều, thiết kế cơ chế hợp nhất đặc trưng dạng thích nghi, tái tổ chức kênh đặc trưng và khai thác tốt hơn các đặc trưng nông cho đối tượng nhỏ.

Một hướng cải tiến tiêu biểu là Dynamic Multi-Path Feature Pyramid Network (DMPFPN), được đề xuất trong mô hình DFA-DETR [15] cho phát hiện đối tượng nhỏ trong ảnh UAV. Khác với FPN truyền thống và cả BiFPN [16], DMPFPN thiết kế luồng thông tin hai chiều với nhiều đường truyền song song (bidirectional multi-path), cho phép tương tác đặc trưng sâu–nông dày đặc hơn ở nhiều mức tỷ lệ. Bên cạnh đó, module Shift-Enhanced Channel Reorganization (SECR) được tích hợp nhằm tái cấu trúc và tăng cường tương tác theo chiều kênh, giúp các kênh mang thông tin chi tiết yếu của đối tượng nhỏ không bị lấn át trong quá trình hợp nhất. Nhờ cơ chế hợp nhất động và tái tổ chức kênh, DMPFPN cải thiện đáng kể khả năng bảo toàn chi tiết không gian của các đối tượng nhỏ. Kết quả thực nghiệm cho thấy DFA-DETR đạt mức cải thiện rõ rệt so với RT-DETR-R18 [11], với mức tăng 3.3% mAP50 và 1.9% mAPs trên VisDrone2019 [1], đồng thời thể hiện tính ổn định cao trên tập HIT-UAV [17], đặc biệt ở chỉ số mAPs.

Song song với hướng tiếp cận dựa trên DETR, một nhánh nghiên cứu khác tập trung vào các bộ phát hiện một giai đoạn nhẹ như YOLO, tiêu biểu là Lightweight-FPN (L-FPN) được đề xuất trong mô hình BPD-YOLO [18]. L-FPN tập trung giải quyết bài toán khoảng cách ngữ nghĩa và chi phí tính toán bằng cách thiết kế Dual-phase Asymptotic Feature Fusion

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

(DAFF) nhằm hợp nhất đặc trưng theo từng giai đoạn, giúp các đặc trưng tầng nông được đưa vào quá trình hợp nhất một cách có kiểm soát. Đồng thời, cơ chế DEI (Decoupled Feature Extraction–Semantic Integration) tách biệt quá trình trích xuất đặc trưng và tích hợp ngữ nghĩa, tránh việc các đặc trưng sâu có độ phân giải thấp làm suy giảm thông tin chi tiết cần thiết cho đối tượng nhỏ. Ngoài ra, module DSPF (Deep Spatial Pyramid Fusion) thay thế các residual blocks nặng bằng chiến lược pooling đa tỷ lệ nhẹ, giúp giảm đáng kể chi phí tính toán trong khi vẫn duy trì khả năng biểu diễn đa tỷ lệ. Thực nghiệm trên VisDrone [1] và TinyPerson [19] cho thấy L-FPN mang lại cải thiện khoảng 2.8% mAP50 so với các baseline YOLO tương ứng, đặc biệt hiệu quả trong các kịch bản UAV mật độ cao và nhiều nhiễu nền.

Tổng hợp lại, các cải tiến FPN gần đây cho thấy một xu hướng rõ ràng: thay vì chỉ mở rộng kiến trúc theo chiều sâu hoặc tăng số tầng pyramid, các nghiên cứu tập trung vào cách thức hợp nhất đặc trưng thông minh hơn, bao gồm luồng thông tin đa đường, cơ chế hợp nhất đặc trưng dạng thích nghi, tái tổ chức kênh và khai thác tối đa đặc trưng tầng nông. DMPFPN đại diện cho hướng tiếp cận thiên về hợp nhất động và tương tác đa tầng mạnh mẽ, phù hợp với kiến trúc DETR và các bài toán yêu cầu ngữ cảnh toàn cục, trong khi L-FPN đại diện cho hướng lightweight và hiệu quả tính toán, phù hợp với các hệ thống UAV thời gian thực. Cả hai đều chứng minh rằng việc thiết kế FPN theo định hướng bài toán cụ thể, đặc biệt là phát hiện đối tượng nhỏ, mang lại cải thiện đáng kể so với FPN và BiFPN truyền thống.

4. ĐÁNH GIÁ THỰC NGHIỆM

4.1. Thiết lập thực nghiệm

4.1.1. Cấu hình thực nghiệm

Các thí nghiệm trên tập VisDrone2019 và TinyPerson được thực hiện trên hệ thống Windows sử dụng GPU NVIDIA RTX 4060 Ti, với môi trường cài đặt Python 3.9 và PyTorch 2.4.0+cu121. Mô hình được huấn luyện bằng bộ tối ưu SGD kết hợp với lịch điều chỉnh tốc độ học hình cosine, với tốc độ học ban đầu 0.01 giảm dần xuống 0.0001 trong 300 epoch, kích thước lô huấn luyện: 8; riêng đối với TinyPerson, kích thước ảnh được đặt là 1024 và kích thước lô giảm xuống 2, trong khi các tham số còn lại được giữ nguyên. Đối với tập HIT-UAV, các thí nghiệm được tiến hành trên hệ thống Ubuntu 20.04 với GPU NVIDIA GeForce RTX 3090 và CPU Intel i9-11900K, sử dụng Python 3.8 và PyTorch 1.13.1. Mô hình được huấn luyện bằng bộ tối ưu AdamW với tốc độ học 0.0001, batch size 8 và 200 epoch.

4.1.2. Bộ dữ liệu đánh giá

Ba bộ dữ liệu được lựa chọn để phản ánh các kịch bản UAV điển hình và bổ sung lẫn nhau, như tóm tắt trong Bảng 1. VisDrone2019 đóng vai trò là bộ dữ liệu chuẩn chính cho bài toán phát hiện đối tượng từ góc nhìn UAV, với tỷ lệ lớn đối tượng nhỏ và sự che khuất phức tạp. TinyPerson đẩy các mô hình đến giới hạn cực đoan khi đối tượng chỉ chiếm vài pixel, phản ánh kịch bản UAV bay ở độ cao lớn. HIT-UAV mở rộng đánh giá sang miền ảnh nhiệt, cho phép kiểm tra khả năng khái quát hóa khi chuyển từ ảnh RGB sang ảnh nhiệt.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Bảng 1. Tổng quan các bộ dữ liệu sử dụng trong đánh giá thực nghiệm.

Bộ dữ liệu	Số ảnh	Số lớp	Đặc trưng chính	Kịch bản
VisDrone2019 [1]	10,209	10	Mật độ cao, nhiều đối tượng nhỏ, che khuất phức tạp	Giám sát đô thị, giao thông
TinyPerson [19]	1,610	1	Đối tượng cực nhỏ	UAV bay cao, quan sát diện rộng
HIT-UAV [17]	2,898	5	Ảnh nhiệt, hạn chế thông tin chi tiết bề mặt	Ban đêm, thời tiết xấu

4.1.3. Chỉ số đánh giá

Các chỉ số được lựa chọn nhằm phản ánh toàn diện cả hiệu suất phát hiện và khả năng triển khai, được liệt kê trong bảng 2.

Bảng 2. Các chỉ số đánh giá được sử dụng trong nghiên cứu

Nhóm chỉ số	Ký hiệu	Ý nghĩa	Đơn vị tính
Độ chính xác	mAP	mAP trung bình tại IoU 0.5:0.95	%
	mAP50	mAP tại IoU = 0.5	%
	mAPs/mAPm/mAPl	AP cho đối tượng nhỏ / trung bình / lớn	%
Hiệu quả	Params	Số tham số mô hình	M (triệu tham số)
	FLOPs	Chi phí tính toán	G (tỷ phép toán)

Trong đó, mAPs đặc biệt quan trọng đối với ứng dụng UAV, nơi phần lớn đối tượng xuất hiện với kích thước nhỏ.

4.2. Đánh giá trên VisDrone2019

Kết quả trên VisDrone2019 được tổng hợp trong Bảng 3, bao gồm các mô hình đại diện cho cả hai hướng tiếp cận dựa trên CNN và dựa trên Transformer. Kết quả cho thấy DFA-DETR đạt mAP50 cao nhất (40.5%) và mAPs cao nhất (14.6%) trên tập VisDrone2019, khẳng định ưu thế rõ rệt của kiến trúc DETR với cơ chế chú ý biến dạng trong việc phát hiện đối tượng nhỏ, vốn là thách thức cốt lõi trong bối cảnh UAV. So với YOLOv11m, DFA-DETR vượt trội cả về mAP50 (+5.5%) và mAPs (+4.8%), đồng thời giảm FLOPs xuống còn khoảng 73% (49.3G so với 67.7G), cho thấy hiệu quả tính toán tốt hơn dù số tham số vẫn ở mức trung bình. Bên cạnh đó, các mô hình BPD-YOLO thể hiện khả năng cân bằng hiệu quả giữa độ chính xác và chi phí tính toán. Đặc biệt, BPD-YOLOn chỉ sử dụng 1.50M tham số và 11.4 GFLOPs nhưng vẫn đạt mAP50 = 38.1%, cho thấy tiềm năng lớn của các thiết kế mô hình nhẹ trong việc khai thác đặc trưng quan trọng mà không cần mở rộng quy mô mạng. Điều này đặc biệt phù hợp cho các hệ thống UAV yêu cầu triển khai thời gian thực và hạn chế tài nguyên.

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

Bảng 3. Kết quả so sánh trên VisDrone2019

Mô hình	mAP50	mAPs	Params	FLOPs
YOLOv11m [15]	35.0	9.8	20.04M	67.7G
YOLOv8s [18]	39.9	–	11.12M	28.5G
DFA-DETR [15]	40.5	14.6	16.55M	49.3G
BPD-YOLOs [18]	45.0	–	5.76M	36.7G
BPD-YOLOn [18]	38.1	–	1.50M	11.4G
RT-DETR-R18 [15]	37.2	12.7	19.89M	57.0G
YOLOv10m [15]	34.5	9.7	15.32M	58.9G

4.3. Đánh giá trên TinyPerson

Kết quả trong Bảng 4 cho thấy YOLOv8n+P2+AFPN đạt mAP@0.5 cao nhất trên tập TinyPerson (39.2%), tuy nhiên đi kèm với chi phí tính toán lớn nhất, với 79.6 GFLOPs và 6.05M tham số. Trong khi đó, BPD-YOLO đạt giá trị mAP@0.5 gần tương đương (39.1%) nhưng chỉ sử dụng 1.50M tham số và 30.9 GFLOPs, thấp nhất trong số các mô hình được so sánh. Điều này cho thấy BPD-YOLO đặc biệt phù hợp cho các ứng dụng UAV yêu cầu hiệu quả tính toán cao và độ tin cậy trong phát hiện đối tượng nhỏ, chẳng hạn như tìm kiếm cứu nạn hoặc giám sát diện rộng. Đáng chú ý, sự chênh lệch mAP giữa các mô hình là rất nhỏ (khoảng 0.1-0.2), trong khi chi phí tính toán khác biệt đáng kể, cho thấy trên tập TinyPerson, nơi các đối tượng có kích thước cực nhỏ, thiết kế kiến trúc và khả năng duy trì bản đồ đặc trưng độ phân giải cao đóng vai trò quan trọng hơn so với việc mở rộng quy mô mạng.

Bảng 4. Kết quả trên TinyPerson

Phương pháp	mAP50	mAP	Params	FLOPs
YOLOv8n+P2+AFPN [18]	39.2	13.8	6.05M	79.6G
YOLOv8n+P2+BiFPN [18]	38.4	13.5	2.34M	49.7G
YOLOv8n+P2 [18]	38.0	13.4	2.93M	31.7G
BPD-YOLO [18]	39.1	13.6	1.50M	30.9G

4.4. Đánh giá trên HIT-UAV

Kết quả trong Bảng 5 cho thấy DFA-DETR đạt hiệu suất cao nhất trên tập HIT-UAV, với mAP50 = 83.1% và mAPs = 43.2%, vượt qua YOLOv11m lần lượt 1.3% và 2.7%. Đồng thời, DFA-DETR cũng đạt AP tổng thể cao nhất (54.8%), cho thấy khả năng phát hiện ổn định trên nhiều thang kích thước đối tượng. Về hiệu quả tính toán, DFA-DETR chỉ sử dụng 49.3 GFLOPs, thấp hơn đáng kể so với YOLOv11m (67.7 GFLOPs), trong khi số tham số cũng giảm từ 20.04M xuống 16.55M. Điều này chứng tỏ thiết kế Transformer với deformable attention không chỉ cải thiện độ chính xác mà còn mang lại lợi thế rõ rệt về chi phí tính toán. Đáng chú ý, trong bối cảnh ảnh nhiệt thiếu thông tin kết cấu và màu sắc, DFA-DETR vẫn duy trì ưu thế rõ ràng, cho thấy khả năng khái quát hóa tốt hơn khi chuyển sang miền dữ liệu mới, đặc biệt phù hợp cho các ứng dụng UAV giám sát và tìm kiếm trong điều kiện ánh sáng kém.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

Bảng 5. Kết quả trên HIT-UAV

Mô hình	mAP50	mAPs	Params	FLOPs
DFA-DETR [15]	83.1	43.2	16.55M	49.3G
YOLOv11m [15]	81.8	40.5	20.04M	67.7G
RT-DETR-R18 [15]	81.1	38.9	19.89M	57.0G

Nhìn chung, kết quả trên ba tập dữ liệu cho thấy hiệu quả của các mô hình không chỉ phụ thuộc vào việc mở rộng quy mô mạng mà còn chịu ảnh hưởng mạnh bởi thiết kế kiến trúc hướng đến phát hiện đối tượng nhỏ. Trên VisDrone2019 và TinyPerson, nơi phần lớn đối tượng có kích thước rất nhỏ và phân bố dày đặc, các mô hình duy trì được bản đồ đặc trưng độ phân giải cao hoặc tận dụng hiệu quả cơ chế chú ý (như deformable attention trong DFA-DETR hoặc các thiết kế gọn nhẹ của BPD-YOLO) cho thấy lợi thế rõ rệt so với các mô hình CNN truyền thống. Mặc dù một số mô hình đạt mAP@0.5 cao hơn, sự chênh lệch về độ chính xác thường khá nhỏ (chỉ khoảng 0.01–0.05), trong khi chi phí tính toán và số tham số có thể khác biệt đáng kể. Điều này cho thấy trong bài toán phát hiện đối tượng nhỏ, hiệu quả trích xuất và duy trì đặc trưng không gian quan trọng hơn so với việc tăng độ sâu hoặc độ rộng của mạng.

Xu hướng này tiếp tục được quan sát trên tập HIT-UAV, nơi ảnh nhiệt thiếu thông tin kết cấu và màu sắc, làm gia tăng độ khó của bài toán. Trong bối cảnh này, DFA-DETR duy trì hiệu suất vượt trội cả về mAP tổng thể và mAPs, cho thấy khả năng khái quát hóa liên miền tốt hơn của kiến trúc Transformer. Ngược lại, các mô hình nhẹ như BPD-YOLO, dù không đạt mAP cao nhất trong mọi trường hợp, lại thể hiện sự cân bằng hợp lý giữa độ chính xác và chi phí tính toán, đặc biệt phù hợp cho các ứng dụng UAV thời gian thực yêu cầu phát hiện ổn định các đối tượng nhỏ với tài nguyên hạn chế.

5. KẾT LUẬN

Bài báo này đã thực hiện một đánh giá thực nghiệm toàn diện các mô hình phát hiện đối tượng hiện đại trong bối cảnh UAV thông qua một khung đánh giá thống nhất trên ba bộ dữ liệu đại diện: VisDrone2019, TinyPerson và HIT-UAV. Kết quả cho thấy hiệu suất phát hiện đối tượng nhỏ đóng vai trò then chốt trong các ứng dụng UAV và không luôn tương quan với mAP tổng thể. Các mô hình quy mô lớn, dù đạt độ chính xác cao, không nhất thiết vượt trội khi xử lý đối tượng nhỏ, trong khi các mô hình nhẹ và các kiến trúc dựa trên Transformer với deformable attention có thể đạt hiệu suất cạnh tranh hoặc vượt trội với chi phí tính toán thấp hơn đáng kể. Đánh giá trên TinyPerson chỉ ra rằng việc mở rộng quy mô mô hình mang lại lợi ích hạn chế khi đối tượng trở nên cực kỳ nhỏ, nhấn mạnh vai trò của thiết kế kiến trúc và xử lý đa tỉ lệ. Đồng thời, các kết quả cross-domain trên HIT-UAV cho thấy các mô hình nhẹ có xu hướng khái quát hóa tốt hơn khi chuyển sang miền dữ liệu mới như ảnh nhiệt. Từ góc độ triển khai, nghiên cứu cũng nhấn mạnh rằng các chỉ số hiệu quả như Params, FLOPs và FPS cần được xem xét đồng thời, do hiệu suất trên phần cứng mạnh không đảm bảo khả năng vận hành thời gian thực trên UAV. Nhìn chung, nghiên cứu này cung cấp một cái nhìn hệ thống về sự đánh đổi giữa độ chính xác, khả năng phát hiện đối tượng nhỏ và chi phí tính toán, đồng thời làm rõ tiềm năng của các kiến trúc nhẹ và dựa trên Transformer hiệu quả trong việc giải quyết các thách thức cốt lõi của bài toán phát hiện đối tượng UAV.

p-ISSN 3030 - 4695
e-ISSN 3030 - 4709

TÀI LIỆU THAM KHẢO

[1] D. Du et al. (2019), *VisDrone-DET2019: The vision meets drone object detection in image challenge results*, in Proceedings of the IEEE/CVF international conference on computer vision workshops, pp.213-226.

[2] T.-Y. Lin et al. (2014), *Microsoft coco: Common objects in context*, in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, pp.740-755.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi (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.

[4] G. J. a. A. C. a. J. Qiu (2023), Ultralytics YOLOv8, Available: https://github.com/ultralytics/ultralytics.

[5] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding (2024), *Yolov10: Real-time end-to-end object detection*, Advances in Neural Information Processing Systems, vol. 37, pp.107984-108011.

[6] R. Khanam and M. Hussain (2024), *Yolov11: An overview of the key architectural enhancements*, arXiv preprint arXiv:2410.17725.

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

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

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

[10] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai (2020), *Deformable detr: Deformable transformers for end-to-end object detection*, arXiv preprint arXiv:2010.04159.

[11] Y. Zhao et al. (2024), *Detrs beat yolos on real-time object detection*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16965-16974.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie (2017), *Feature pyramid networks for object detection*, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp.2117-2125.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick (2017), *Mask r-cnn*, in Proceedings of the IEEE international conference on computer vision, pp.2961-2969.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár (2017), *Focal loss for dense object detection*, in Proceedings of the IEEE international conference on computer vision, pp.2980-2988.

[15] B. Zhang and Y. Zhang (2025), *UAV Small Object Detection Algorithm Based on Dynamic Feature Aggregation and Hierarchical Attention Mechanism*, IEEE Access.

p-ISSN 3030 - 4628
e-ISSN 3030 - 4636

[16] M. Tan, R. Pang, and Q. V. Le (2020), *Efficientdet: Scalable and efficient object detection*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.10781-10790.

[17] J. Suo, T. Wang, X. Zhang, H. Chen, W. Zhou, and W. Shi (2023), *HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection*, Scientific Data, 10(1), pp.227.

[18] M. Chao, C. Peng, L. Yun, C. Zhang, H. Wang, and Z. Chen (2025), *A lightweight small object detection model for UAV images based on deep semantic integration*, Scientific Reports, 15(1), pp. 31888.

[19] X. Yu, Y. Gong, N. Jiang, Q. Ye, and Z. Han (2020), *Scale match for tiny person detection*, in Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp.1257-1265.

EVALUATING THE EFFECTIVENESS OF FEATURE FUSION–BASED OBJECT DETECTION MODELS IN UAV IMAGERY

Nguyen Dung, Nguyen Ngoc Thuy, Bui Luong Vu Ngoc

ABSTRACT

Object detection from the UAV perspective has attracted increasing attention due to its importance in applications such as traffic monitoring, smart agriculture, and environmental observation. However, UAV imagery often contains small, densely distributed objects with frequent occlusions and complex backgrounds, posing significant challenges. This paper conducts a comprehensive survey and experimental evaluation of modern object detection models based on CNNs and Transformers in UAV scenarios, using the VisDrone2019, TinyPerson, and HIT-UAV benchmarks. The results reveal a clear trade-off between detection accuracy and computational cost, while recent approaches such as adaptive attention mechanisms and multi-scale feature pyramid architectures demonstrate strong potential for achieving a favorable balance between performance and efficiency in UAV deployment.

Keywords: *UAV, object detection, deep learning, attention mechanism, multi-level feature pyramid.*

** Ngày nộp bài:26/01/2026; Ngày gửi phản biện:27/01/2026; Ngày duyệt đăng: 28/02/2026*