ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC

Phát Hiện Đối Tượng Trên Ảnh Hồng Ngoại Tầm Cao

Nhóm 15

Báo cáo giữa kỳ
Một số vấn đề chọn lọc về thị giác máy tính

Thành viên:
Nguyễn Thừa Tuân - 22001652
Nguyễn Hoàng Việt - 22001658
Nguyễn Quang Việt - 22001659

Giảng viên:
TS Cao Văn Chung

Hà Nội, Ngày 4 tháng 1 năm 2026

Chương 1

Lời cảm ơn

Trước tiên, chúng em xin gửi lời cảm ơn chân thành đến TS. Cao Văn Chung, người đã tận tâm giảng dạy học phần Chọn lọc thị giác máy tính, một môn học nâng cao dựa trên nền tảng kiến thức môn Thị giác máy tính mà chúng em đã được học ở học kỳ trước. Những kiến thức lý thuyết và thực hành chuyên sâu do thầy truyền đạt đã tạo cơ sở vững chắc để nhóm triển khai đề tài "Phát hiện đối tượng trên ảnh hồng ngoại tầm cao".

Bên cạnh đó, những định hướng, góp ý và sự hỗ trợ kịp thời của thầy trong suốt quá trình thực hiện đề tài đã giúp nhóm nhận diện rõ các hạn chế còn tồn tại, đồng thời có phương án điều chỉnh phù hợp để nâng cao chất lượng mô hình. Sự tận tâm và trách nhiệm của thầy là nguồn động lực to lớn giúp nhóm hoàn thành đề tài này.

Nhóm cũng trân trọng cảm ơn các học phần đã được học tại trường, đặc biệt là các môn thị giác máy tính, học máy. Những kiến thức này đã giúp nhóm có nền tảng để tổ chức, phân tích, thiết kế mô hình, xây dựng quy trình thực nghiệm và trình bày kết quả một cách khoa học.

Trong quá trình thực hiện, nhóm cũng gặp phải không ít khó khăn kỹ thuật, đặc biệt là trong việc tinh chỉnh mô hình và tối ưu hóa hiệu năng trên dữ liệu ảnh hồng ngoại. Tuy nhiên, nhờ những kiến thức nền tảng và sự định hướng từ thầy, cùng với tinh thần tìm tòi và trao đổi nội bộ trong nhóm, chúng em đã từng bước khắc phục các vấn đề và hoàn thiện mô hình đúng theo mục tiêu đề tài. Chúng em xin bày tỏ lòng biết ơn sâu sắc đến thầy vì sự tận tâm và hỗ trợ trong suốt quá trình này.

Cuối cùng, xin gửi lời cảm ơn chân thành đến tất cả các thành viên trong nhóm. Sự hợp tác nghiêm túc, tinh thần trách nhiệm và những đóng góp ý tưởng xuyên suốt quá trình làm việc đã giúp đề tài được triển khai một cách hiệu quả và đạt được kết quả như kỳ vọng.

Một lần nữa, chúng em xin trân trọng cảm ơn.

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG

Chương 2

Giới thiệu

2.1 Đặt vấn đề

Trong những năm gần đây, sự phát triển mạnh mẽ của lĩnh vực thị giác máy tính và trí tuệ nhân tạo đã đem lại nhiều bước tiến quan trọng trong việc tự động hóa các tác vụ phân tích hình ảnh. Các mô hình hiện đại như mạng nơ-ron tích chập, họ mô hình YOLO, Faster R-CNN hay các mô hình dựa trên Transformer đã đạt được độ chính xác cao trong nhiều bài toán như phân loại ảnh, phát hiện đối tượng, theo dõi mục tiêu và phân đoạn ảnh. Tuy nhiên, phần lớn các nghiên cứu nổi bật hiện nay đều tập trung vào ảnh ánh sáng khả kiến (RGB), trong khi ảnh hồng ngoại – đặc biệt là ảnh hồng ngoại thu từ độ cao lớn – vẫn còn tương đối ít được khai thác trong cộng đồng nghiên cứu.

Ảnh hồng ngoại tầm cao đóng vai trò quan trọng trong nhiều ứng dụng thực tiễn: giám sát đường không, cảnh báo cháy rừng, tìm kiếm cứu nạn, quan sát biên giới, giám sát hoạt động hàng không dân sự và quân sự, cũng như trong các hệ thống dẫn đường hay cảnh giới hiện đại. Loại dữ liệu này có ưu điểm vượt trội khi hoạt động tốt trong điều kiện thiếu sáng, sương mù, hoặc ban đêm, nơi các camera RGB truyền thống gặp nhiều hạn chế. Tuy nhiên, việc phát hiện đối tượng trong ảnh hồng ngoại tầm cao vẫn là một thách thức lớn.

Thứ nhất, ảnh hồng ngoại thường có độ tương phản thấp, chi tiết mờ và cấu trúc hình dạng của đối tượng không rõ ràng. Các đối tượng như máy bay, UAV, phương tiện di chuyển hoặc con người có thể bị hòa lẫn vào nền nhiệt của môi trường, đặc biệt trong điều kiện thời

tiết thay đổi hoặc nhiệt độ môi trường cao.

Thứ hai, góc quan sát từ trên cao khiến kích thước đối tượng trong ảnh rất nhỏ (small-object detection), đôi khi chỉ chiếm vài pixel. Đây là một trong những bài toán khó nhất của phát hiện đối tượng, bởi phần lớn các mô hình thị giác máy tính đều suy giảm hiệu suất đáng kể khi làm việc với mục tiêu kích thước nhỏ.

Thứ ba, dữ liệu hồng ngoại tầm cao rất khó thu thập và thường mang tính đặc thù cao. Phần lớn những bộ dữ liệu hiện có liên quan đến quan sát quân sự hoặc an ninh nên không được công bố rộng rãi. Điều này gây hạn chế cho việc huấn luyện, đánh giá và so sánh mô hình. Một mô hình tốt yêu cầu dữ liệu đa dạng, có chất lượng và được gán nhãn chính xác – song đây lại là điểm yếu lớn trong lĩnh vực ảnh hồng ngoại.

Những thách thức nêu trên tạo ra khoảng trống nghiên cứu đáng kể trong việc phát triển các mô hình phát hiện đối tượng trên ảnh hồng ngoại tầm cao. Việc nghiên cứu và đề xuất một phương pháp hiệu quả cho bài toán này không chỉ mang ý nghĩa lý thuyết mà còn góp phần quan trọng vào các ứng dụng giám sát, an ninh và hàng không.

Xuất phát từ bối cảnh đó, đề tài "Phát hiện đối tượng trên ảnh hồng ngoại tầm cao" được thực hiện với mục tiêu xây dựng một quy trình hoàn chỉnh từ tiền xử lý dữ liệu, lựa chọn mô hình phù hợp, tối ưu hóa tham số, huấn luyện và đánh giá mô hình, nhằm đưa ra một giải pháp có tính khả thi và hiệu quả cao. Đồng thời, đề tài hướng tới việc xây dựng hoặc mở rộng một bộ dữ liệu hồng ngoại tầm cao chất lượng, tạo nền tảng cho các nghiên cứu tiếp theo trong lĩnh vực này.

2.2 Mục tiêu

Mục tiêu tổng quát của đề tài là nghiên cứu và xây dựng một mô hình phát hiện đối tượng hoạt động hiệu quả trên ảnh hồng ngoại tầm cao. Từ mục tiêu tổng quát đó, các mục tiêu cụ thể bao gồm:

- **Khảo sát và phân tích đặc trưng của ảnh hồng ngoại tầm cao:** Tìm hiểu các yếu tố ảnh hưởng đến chất lượng ảnh, độ phân giải, độ tương phản và các dạng nhiễu

phổ biến.

- **Xây dựng hoặc chuẩn hóa bộ dữ liệu hồng ngoại tầm cao:** Thu thập, chọn lọc, gán nhãn và tổ chức dữ liệu theo cấu trúc phù hợp cho huấn luyện mô hình học sâu.
- **Nghiên cứu và lựa chọn mô hình phát hiện đối tượng phù hợp:** Thử nghiệm các mô hình hiện đại như YOLO để tìm ra phương pháp hiệu quả nhất cho dữ liệu hồng ngoại.
- **Huấn luyện, tối ưu hóa và đánh giá mô hình:** Tiến hành huấn luyện mô hình trên bộ dữ liệu hồng ngoại tầm cao, tinh chỉnh siêu tham số và đánh giá dựa trên các chỉ số như mAP, Precision, Recall.
- **Đề xuất giải pháp cải thiện hiệu suất:** Áp dụng các kỹ thuật như tăng cường dữ liệu, nâng cao độ phân giải, module tăng cường thông tin biên, hoặc các phương pháp xử lý đặc thù cho small-object detection.
- **Xây dựng báo cáo và đề xuất hướng phát triển tiếp theo:** Tổng hợp kết quả, phân tích hạn chế và đề xuất các hướng mở cho các nghiên cứu sau.

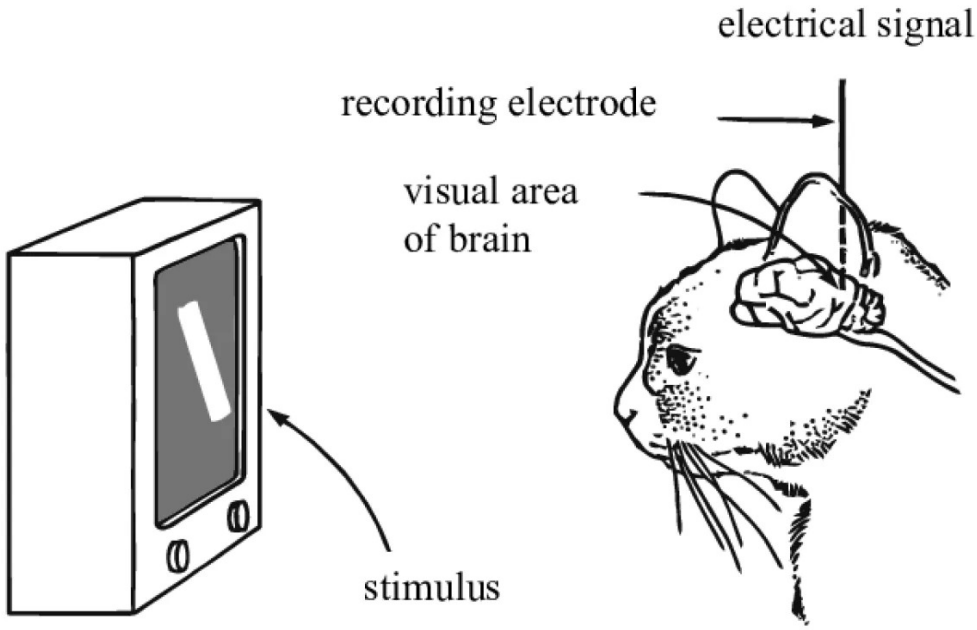
Thông qua các mục tiêu trên, đề tài hướng đến việc đóng góp một hướng tiếp cận thực tiễn và hiệu quả cho bài toán phát hiện đối tượng trong ảnh hồng ngoại tầm cao – một lĩnh vực còn nhiều thách thức nhưng có ý nghĩa quan trọng trong giám sát và định vị mục tiêu từ nền tảng trên không.

2.3 Bối cảnh: Từ Thị giác Máy tính đến nhu cầu phát hiện Thời gian thực

Để hiểu được tầm quan trọng của YOLO, trước tiên chúng ta phải nhìn lại lịch sử của thị giác máy tính và những thách thức mà các nhà nghiên cứu phải đối mặt.

2.3.1 Nguồn gốc của Thị giác Máy tính

Thị giác máy tính, một lĩnh vực con của trí tuệ nhân tạo, tập trung vào việc dạy máy móc "nhìn" và hiểu thế giới trực quan. Nguồn gốc của nó bắt nguồn từ cuối những năm 1950 và đầu những năm 1960. Các nghiên cứu tiên phong của David Hubel và Torsten Wiesel về cách não mèo phản ứng với các mẫu đơn giản như cạnh và đường thẳng đã đặt nền móng cho khái niệm "trích xuất đặc trưng" . Song song đó, sự phát triển của công nghệ cho phép



Hình 2.1: Nghiên cứu cách não mèo phản ứng với các thanh ánh sáng đã giúp phát triển trích xuất đặc trưng trong thị giác máy tính.

biến hình ảnh vật lý thành định dạng kỹ thuật số đã khơi dậy sự quan tâm lớn. "Summer Vision Project" của MIT vào năm 1966, dù không thành công hoàn toàn, đã đánh dấu sự khởi đầu chính thức của thị giác máy tính như một lĩnh vực khoa học, với mục tiêu tách tiền cảnh khỏi hậu cảnh.

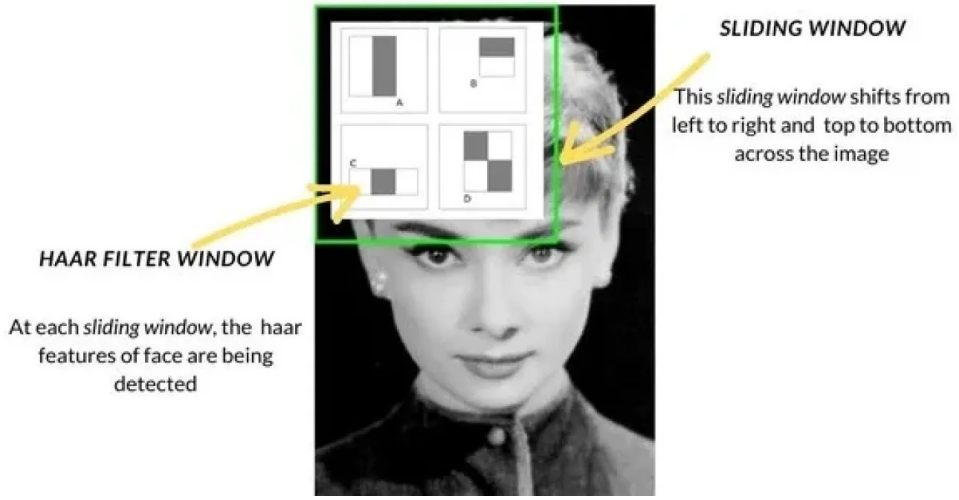
2.3.2 Các phương pháp Phát hiện Đối tượng Sơ khai

Khi lĩnh vực này phát triển, các phương pháp phát hiện đối tượng đã chuyển từ kỹ thuật cơ bản như "so khớp mẫu" sang các cách tiếp cận tiên tiến hơn:

Haar Cascade: Trở nên phổ biến vào đầu những năm 2000, đặc biệt trong nhận diện khuôn

mặt. Phương pháp này hoạt động bằng cách quét một cửa sổ trượt (sliding window) trên hình ảnh, kiểm tra các đặc trưng cụ thể (như cạnh, kết cấu) và kết hợp chúng. Mặc dù nhanh hơn các phương pháp trước đó, nó vẫn gặp khó khăn với sự thay đổi về góc nhìn và ánh sáng.

HOG và SVM: Biểu đồ các Gradient Định hướng (Histogram of Oriented Gradients - HOG) được giới thiệu, cũng sử dụng kỹ thuật cửa sổ trượt để phân tích sự thay đổi ánh sáng và bóng tối, giúp xác định đối tượng dựa trên hình dạng. Sau đó, Máy vectơ hỗ trợ (Support Vector Machines - SVM) được sử dụng để phân loại các đặc trưng này. Các phương pháp này cải thiện độ chính xác nhưng vẫn còn chậm và thiếu ổn định trong môi trường thực tế.



Hình 2.2: Sử dụng Haar Cascade để Nhận diện khuôn mặt.

2.3.3 Deep Learning và Giới hạn của R-CNN

Những năm 2010 chứng kiến sự bùng nổ của học sâu (Deep Learning) và Mạng nơ-ron tích chập (Convolutional Neural Networks - CNN). CNN cho phép máy tính tự động học các đặc trưng quan trọng từ dữ liệu, dẫn đến độ chính xác tăng vọt.

Các mô hình như R-CNN (Mạng nơ-ron tích chập dựa trên vùng) là một cải tiến lớn, nhưng chúng hoạt động qua nhiều giai đoạn:

1. Đề xuất các vùng có khả năng chứa đối tượng.

2. Trích xuất đặc trưng CNN cho từng vùng.

3. Phân loại các vùng đó.

Quá trình nhiều bước này khiến R-CNN và các biến thể sau đó như Fast R-CNN và Faster R-CNN, tuy chính xác hơn, nhưng vẫn quá chậm, không thể đáp ứng nhu cầu ngày càng tăng về phát hiện đối tượng theo thời gian thực (real-time detection) trong các ứng dụng như xe tự lái hay giám sát video.

Bảng 2.1: So sánh thời gian xử lý giữa R-CNN, Fast R-CNN và Faster R-CNN

	R-CNN	Fast R-CNN	Faster R-CNN
Test Time per Image	50 Seconds	2 Seconds	0.2 Seconds
Speed Up	1x	25x	250x

2.4 YOLO (You Only Look Once): Một Cột mốc Quan trọng

Chính trong bối cảnh cần bằng giữa tốc độ và độ chính xác, YOLO (You Only Look Once) đã xuất hiện và định nghĩa lại hoàn toàn bài toán phát hiện đối tượng.

2.4.1 Triết lý và Kiến trúc cốt lõi

Không giống như các hệ thống R-CNN đa giai đoạn, kiến trúc của YOLO coi việc phát hiện đối tượng như một bài toán hồi quy (regression problem) duy nhất. Thay vì phân tích từng vùng đề xuất, YOLO xử lý toàn bộ hình ảnh chỉ trong một lần duy nhất:

1. **Chia lưới (Grid System):** Hình ảnh đầu vào được chia thành một lưới (ví dụ: $S \times S$).
2. **Dự đoán đồng thời:** Mỗi grid cell chịu trách nhiệm dự đoán các đối tượng có tâm rơi vào ô đó.
3. **Đầu ra hợp nhất:** Mỗi ô dự đoán B bounding boxes cùng với confidence score cho mỗi box, và C xác suất lớp cho đối tượng trong ô đó.

Kết quả là một tensor dự đoán duy nhất, bao gồm cả vị trí và phân loại lớp, được tạo ra chỉ sau một lượt truyền thuận (single forward pass) qua mạng nơ-ron.

2.4.2 Tác động và Ưu điểm

Cách tiếp cận đột phá này mang lại những lợi ích to lớn:

- **Tốc độ thời gian thực:** Do kiến trúc một giai đoạn, YOLO cực kỳ nhanh, có khả năng xử lý video theo thời gian thực (ví dụ: 45 FPS hoặc cao hơn trên các GPU thời đó).
- **Hiểu biết toàn cục:** YOLO nhìn toàn bộ hình ảnh trong quá trình dự đoán, do đó nó nắm bắt được ngữ cảnh toàn cục, giúp giảm đáng kể lỗi dự đoán sai hậu cảnh (false

positives) so với các phương pháp dựa trên cửa sổ trượt.

Do tốc độ, độ chính xác và bản chất mã nguồn mở, YOLO nhanh chóng trở thành một lựa chọn hàng đầu cho các giải pháp thời gian thực trong nhiều ngành công nghiệp, từ sản xuất, chăm sóc sức khỏe đến robot học.

2.5 Lịch sử phát triển của các mô hình YOLO

Từ nền tảng YOLO ban đầu, cộng đồng mã nguồn mở và các tổ chức như Ultralytics đã liên tục cải tiến, xây dựng dựa trên những tiến bộ của mỗi phiên bản. Trọng tâm không chỉ là cải thiện hiệu suất mà còn là làm cho các mô hình này dễ tiếp cận và dễ sử dụng hơn.

2.5.1 Ultralytics YOLOv5

Khi Ultralytics giới thiệu YOLOv5, một trong những đóng góp lớn nhất là việc đơn giản hóa việc triển khai và sử dụng.

- **Nền tảng PyTorch:** Được xây dựng trên PyTorch, một framework học sâu phổ biến, YOLOv5 giúp các nhà phát triển dễ dàng tích hợp, đào tạo và gỡ lỗi hơn.
- **Tập trung vào khả năng sử dụng:** Nó kết hợp giữa độ chính xác cao và khả năng sử dụng thực tế, cho phép nhiều người dùng hơn, ngay cả những người không phải là chuyên gia về coding, có thể triển khai các giải pháp phát hiện đối tượng tiên tiến.

YOLOv5 đã trở thành một trong những mô hình phát hiện đối tượng phổ biến nhất, tạo nền tảng vững chắc cho các thế hệ tiếp theo.

2.5.2 Ultralytics YOLOv8

Ultralytics YOLOv8 tiếp tục sự tiến bộ này bằng cách không chỉ cải thiện mô hình phát hiện mà còn mở rộng đáng kể các khả năng của nó. YOLOv8 được thiết kế như một framework thống nhất, hỗ trợ nhiều tác vụ thị giác máy tính, bao gồm:

- Phát hiện đối tượng
- Phân đoạn thể hiện
- Phân loại hình ảnh
- Ước tính tư thế

Sự linh hoạt này làm cho YOLOv8 trở thành một công cụ mạnh mẽ, có thể được sử dụng cho cả các ứng dụng cơ bản lẫn các hệ thống phức tạp, đòi hỏi nhiều loại phân tích thị giác khác nhau.

Chương 3

Mô tả dữ liệu

3.1 Giới thiệu chung

Bộ dữ liệu HIT-UAV là một tập dữ liệu ảnh hồng ngoại tầm cao được xây dựng nhằm giải quyết nhu cầu cấp thiết trong lĩnh vực phát hiện đối tượng từ các phương tiện bay không người lái (UAV). Khác với các bộ dữ liệu truyền thống vốn chủ yếu bao gồm ảnh RGB thu từ mặt đất hoặc từ độ cao thấp, HIT-UAV tập trung vào dạng dữ liệu hồng ngoại nhiệt (thermal infrared) được thu thập từ UAV bay ở độ cao lớn trong các môi trường thực tế, bao gồm trường học, bãi đỗ xe, đường giao thông và sân chơi. Tổng cộng 2.898 ảnh hồng ngoại đã được lựa chọn từ 43.470 khung hình video gốc, đảm bảo giữ lại các khung hình đa dạng, có sự thay đổi đủ lớn và chứa nhiều đối tượng trong cùng cảnh.

Điểm đặc biệt của HIT-UAV không chỉ nằm ở dạng ảnh hồng ngoại mà còn ở việc bộ dữ liệu cung cấp toàn bộ thông tin liên quan đến chuyến bay cho từng ảnh, bao gồm độ cao bay, góc nghiêng của camera, cường độ ánh sáng môi trường, thời gian chụp (ngày hoặc đêm) và điều kiện thời tiết. Đây là những thông tin mà rất nhiều bộ dữ liệu UAV trước đây không ghi nhận, khiến chúng khó được sử dụng trong các nghiên cứu chuyên sâu về tác động của góc nhìn và độ cao đối với khả năng phát hiện đối tượng.

Một mục tiêu quan trọng của HIT-UAV là giải quyết vấn đề thiếu hụt các bộ dữ liệu hồng ngoại ở độ cao lớn dành cho phát hiện người và phương tiện. Trong khi các bộ dữ liệu như VisDrone, UAV123 hay CARPK chủ yếu tập trung vào ảnh RGB, và các bộ dữ liệu ảnh

nhiệt khác như ASL-TID, BIRDSAI hay FLAME lại chỉ thu thập ở độ cao thấp hoặc trong các bối cảnh hạn chế, HIT-UAV mang lại một nguồn dữ liệu thực nghiệm có độ đa dạng lớn hơn nhiều, đặc biệt phù hợp với các kịch bản thực tế đòi hỏi UAV hoạt động trong môi trường thiếu sáng hoặc hoàn toàn không có ánh sáng.

3.2 Nền tảng thu thập dữ liệu và thiết bị cảm biến

Quá trình thu thập dữ liệu của HIT-UAV được thực hiện bằng UAV DJI Matrice M210 V2, một dòng UAV công nghiệp được thiết kế cho các nhiệm vụ quan sát tầm xa, an ninh, cứu hộ và thu thập dữ liệu địa không gian. UAV này có kích thước lớn, trọng tâm ổn định và khả năng chịu tải cao, giúp nó mang được cảm biến hồng ngoại chuyên dụng trong thời gian bay đủ dài để thu thập dữ liệu có chất lượng tốt. Khi cất cánh với camera Zenmuse XT2, UAV duy trì trọng lượng tổng thể khoảng 6.14 kg và thời gian bay tối đa từ 24 đến 34 phút tuỳ tải trọng.

Camera chính sử dụng để ghi nhận dữ liệu là DJI Zenmuse XT2. Đây là thiết bị kết hợp giữa camera RGB và camera hồng ngoại FLIR, trong đó phần cảm biến nhiệt có độ phân giải 640×512 pixel và sử dụng ống kính 25 mm. Do hoạt động trong dải bước sóng dài của bức xạ hồng ngoại, camera có khả năng ghi lại sự khác biệt nhiệt độ của các vật thể, ngay cả trong điều kiện hoàn toàn không có ánh sáng khả kiến. Điều này cho phép thu được các ảnh mà trong đó người và phương tiện thường nổi bật so với nền do sự chênh lệch nhiệt độ, đặc biệt là vào ban đêm.

3.3 Quy trình thu thập và tạo lập bộ dữ liệu

Quy trình tạo lập bộ dữ liệu bao gồm bốn bước: ghi hình video, trích xuất khung hình và làm sạch dữ liệu, gán nhãn đối tượng, và cuối cùng là sinh tập dữ liệu hoàn chỉnh theo định dạng chuẩn.

3.3.1 Ghi hình video

Việc ghi hình được thực hiện trong nhiều môi trường thực, bao gồm khuôn viên trường đại học, bãi đỗ xe, đường phố, khu dân cư và sân chơi. Mục tiêu của nhóm tác giả là ghi lại các cảnh có đông người và phương tiện, hoặc các cảnh có mật độ đối tượng thay đổi tùy thời điểm. Độ cao bay được điều chỉnh từ 60 đến 130 m, với các bước 10 m để thu thập dữ liệu có sự thay đổi có hệ thống theo độ cao. Tương tự, góc camera được điều chỉnh từ 30° đến 90° để quan sát tác động của góc nhìn lên kích thước và hình dạng của đối tượng. Tất cả các video đều được quay với tốc độ 7 FPS và đảm bảo điều kiện không mưa nhằm tránh hiện tượng nhiễu do hơi nước hoặc giọt nước làm giảm chất lượng hình ảnh.

Quá trình ghi hình diễn ra cả ban ngày và ban đêm. Điều này rất quan trọng vì ảnh nhiệt ban đêm thường có độ tương phản cao hơn do chênh lệch nhiệt độ giữa vật thể và nền lớn hơn, trong khi ảnh ban ngày có độ tương phản thấp hơn do bề mặt môi trường hấp thụ và tỏa nhiệt mạnh. Sự khác biệt giữa ảnh ngày và đêm giúp mô hình học được các đặc trưng đa dạng và xử lý tốt hơn các tình huống ngoài thực tế.

3.3.2 Trích xuất khung hình và xử lý dữ liệu

Từ tổng cộng 43.470 khung hình, nhóm tác giả tiến hành trích xuất một ảnh mỗi 15 khung nhằm loại bỏ các khung hình gần như giống nhau. Vì tốc độ quay chỉ 7 FPS, các khung hình liên tiếp không mang lại sự khác biệt đủ lớn để cải thiện quá trình huấn luyện mô hình. Do đó, lựa chọn khoảng cách 15 khung hình giúp giữ lại các thay đổi về vị trí, hình dạng và vị trí tương đối của đối tượng, đồng thời giảm dung lượng lưu trữ mà không ảnh hưởng đến chất lượng dữ liệu.

Kết quả của bước này là bộ 2.898 ảnh nhiệt chất lượng cao, có sự đa dạng lớn về bố cục cảnh, mật độ đối tượng, điều kiện ánh sáng và góc nhìn.

3.3.3 Gán nhãn đối tượng

Giai đoạn gán nhãn là một phần quan trọng trong việc xây dựng HIT-UAV. Nhóm tác giả sử dụng một phiên bản sửa đổi của công cụ LabelImg để hỗ trợ gán nhãn cả hai loại bounding box: bounding box chuẩn và bounding box xoay. Các đối tượng được ghi nhận thuộc một trong bốn loại chính gồm người, ô tô, xe đạp và các phương tiện khác (như xe tải nhỏ hoặc phương tiện đặc thù). Ngoài ra, một số vùng ảnh không thể phân biệt rõ ràng loại đối tượng do kích thước quá nhỏ hoặc bị che khuất được gán vào lớp *DontCare* nhằm tránh gây nhiễu cho mô hình trong quá trình huấn luyện.

Bounding box xoay đặc biệt hữu ích trong các tình huống mà các phương tiện hoặc người được quan sát từ góc nghiêng lớn, khiến hình dạng của chúng không còn thẳng đứng trong khung hình. So với bounding box chuẩn, bounding box xoay giảm đáng kể mức độ chồng lấn giữa các đối tượng trong các cảnh đông đúc. Tuy nhiên, vì không phải mô hình nào cũng hỗ trợ bounding box xoay, nhóm tác giả cung cấp cả hai phiên bản để người dùng linh hoạt lựa chọn.

Mỗi bounding box xoay ban đầu được chuyển đổi sang bounding box chuẩn bằng cách tìm giá trị nhỏ nhất và lớn nhất theo trục toạ độ, qua đó xác định hình chữ nhật bao ngoài. Toàn bộ quá trình gán nhãn được thực hiện bởi ba người độc lập và có kiểm chứng chéo nhằm đảm bảo tính nhất quán và độ chính xác cao.

3.3.4 Sinh tập dữ liệu hoàn chỉnh

Sau khi gán nhãn, nhóm tác giả phát triển một công cụ sinh dữ liệu tự động để xuất các file nhãn theo hai chuẩn phổ biến: VOC XML và COCO JSON. Bên cạnh đó, mỗi dạng nhãn đều có phiên bản cho bounding box chuẩn và bounding box xoay. Bộ dữ liệu cuối cùng gồm bốn thư mục: *normal_xml*, *normal_json*, *rotate_xml* và *rotate_json*.

Dữ liệu được chia theo tỉ lệ 70% cho tập huấn luyện, 20% cho tập kiểm thử và 10% cho tập kiểm định. Tỉ lệ này được lựa chọn nhằm đảm bảo mỗi phân đoạn đều có số lượng đối tượng và bối cảnh đa dạng, từ đó giúp mô hình huấn luyện có khả năng tổng quát tốt hơn.

3.4 Phân bố dữ liệu và đặc điểm thống kê

Toàn bộ bộ dữ liệu bao gồm 24.899 đối tượng được gán nhãn, trong đó lớp người chiếm tỷ lệ lớn nhất do đặc thù của các bối cảnh thu thập. Số lượng ô tô và xe đạp cũng xuất hiện nhiều trong các khu vực bãi đỗ xe và đường giao thông, giúp dữ liệu phản ánh đúng đặc điểm môi trường đô thị. Lớp phương tiện khác xuất hiện ít hơn, dẫn đến sự mất cân bằng dữ liệu nhất định, điều mà nhóm tác giả cũng đã chỉ ra trong bài báo khi phân tích kết quả thực nghiệm.

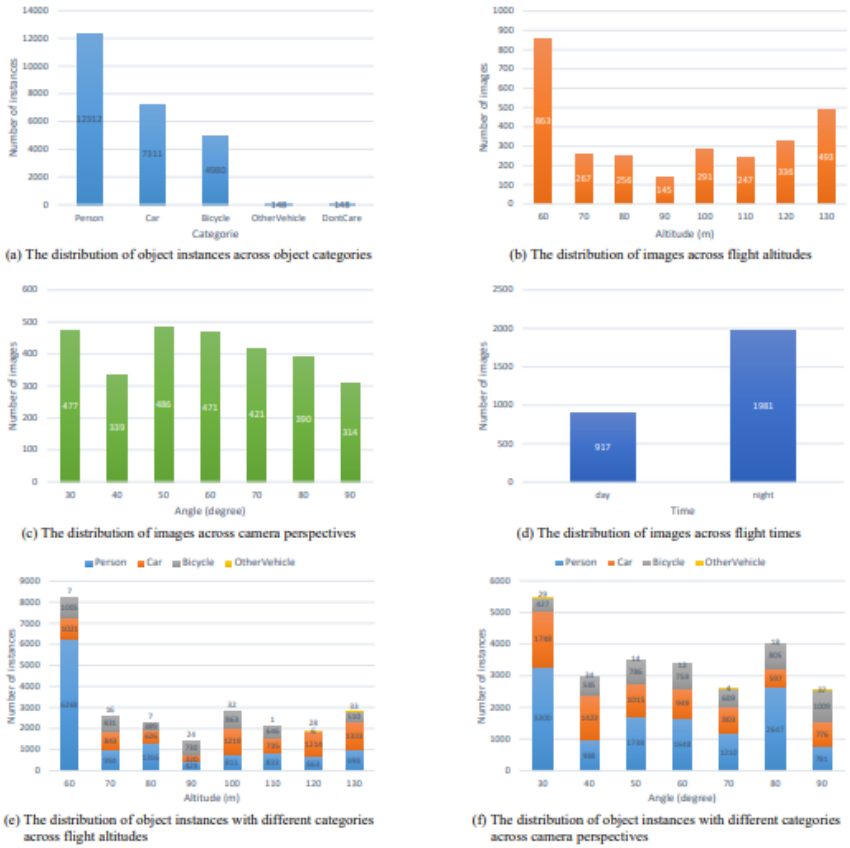
Phân bố dữ liệu theo độ cao cho thấy số lượng ảnh thu được tại các độ cao 60 m, 70 m và 80 m cao hơn so với các độ cao lớn hơn. Điều này phản ánh đặc điểm vận hành UAV, khi các độ cao trung bình cho phép ghi nhận nhiều đối tượng hơn trong cùng một khung hình mà vẫn đảm bảo độ phân giải đủ lớn để gán nhãn. Trong khi đó, góc camera thay đổi từ 30° đến 90° cho thấy sự đa dạng về góc nhìn. Ở góc 30°, tầm quan sát rộng và đối tượng ở xa thường nhỏ, trong khi góc 90° cho phép quan sát thẳng đứng nhưng làm giảm diện tích bề mặt nhìn thấy của đối tượng. Dữ liệu phân bố đều ở các góc nhìn giúp phân tích đầy đủ tác động của từng góc đến khả năng phát hiện.

Bên cạnh đó, sự khác biệt giữa ảnh ngày và đêm cũng là một yếu tố quan trọng của HIT-UAV. Theo phân tích trong bài báo, ảnh nhiệt ban đêm có độ tương phản cao hơn và giúp mô hình đạt độ chính xác tốt hơn. Việc đưa cả hai loại ảnh vào bộ dữ liệu làm tăng độ đa dạng và giúp mô hình học được các đặc trưng phức tạp hơn liên quan đến sự thay đổi nhiệt độ bề mặt môi trường.

3.5 Cấu trúc tên file và tổ chức thư mục

Mỗi ảnh trong bộ dữ liệu được đặt tên theo cấu trúc mã hoá chứa thông tin về thời gian, độ cao, góc camera, điều kiện thời tiết và số thứ tự. Cách mã hoá này đảm bảo người sử dụng có thể nhanh chóng truy xuất đúng nhóm ảnh phục vụ phân tích, ví dụ như nghiên cứu ảnh hưởng của độ cao 100 m ở góc camera 50° vào thời điểm ban đêm.

Bốn thư mục dữ liệu chính được tổ chức theo hai trục: loại bounding box và định dạng



Hình 3.1: Phân bố dữ liệu

nhãn. Việc cung cấp cả hai định dạng VOC và COCO giúp bộ dữ liệu tương thích với hầu hết các mô hình phát hiện đối tượng phổ biến hiện nay như YOLOv4, Faster R-CNN, SSD hoặc các mô hình Transformer-based.

3.6 Ý nghĩa và tầm quan trọng của bộ dữ liệu

HIT-UAV là một trong những bộ dữ liệu quan trọng nhất hiện nay dành cho nghiên cứu phát hiện đối tượng trong ảnh hồng ngoại tầm cao từ UAV. Với độ đa dạng lớn về bối cảnh, độ cao, góc nhìn và điều kiện ánh sáng, bộ dữ liệu này cung cấp một nền tảng vững chắc để đánh giá khả năng tổng quát của các mô hình học sâu. Việc bổ sung đầy đủ thông tin bay và hai dạng nhãn (chuẩn và xoay) giúp mở ra nhiều hướng nghiên cứu mới, bao gồm phân tích ảnh hưởng của góc nhìn và độ cao, tối ưu quỹ đạo bay, hoặc cải thiện khả năng phát hiện đối tượng nhỏ trong các cảnh phức tạp.

Đặc biệt, kết quả thực nghiệm trong bài báo cho thấy các mô hình truyền thống như

YOLOv4, SSD hay Faster R-CNN đều đạt độ chính xác rất cao trên HIT-UAV, vượt xa kết quả trên các bộ dữ liệu RGB như COCO hoặc VisDrone. Điều này chứng minh hiệu quả vượt trội của ảnh nhiệt trong các nhiệm vụ giám sát, tìm kiếm cứu nạn, an ninh và nhận dạng phương tiện vào ban đêm hoặc điều kiện ánh sáng yếu.

Từ những phân tích trên, có thể thấy rằng HIT-UAV không chỉ là một bộ dữ liệu tạo ra để lấp đầy khoảng trống của các bộ dữ liệu ảnh nhiệt tầm cao, mà còn là nền tảng nghiên cứu hoàn chỉnh, hỗ trợ cả khoa học cơ bản lẫn ứng dụng thực tế trong lĩnh vực UAV và thị giác máy tính.

Chương 4

Phương pháp thực hiện

4.1 Tổng quan hệ thống theo dõi đa đối tượng

Hệ thống theo dõi đa đối tượng đề xuất được thiết kế theo kiến trúc *Tracking-by-Detection* . Đây là phương pháp tiêu chuẩn cho các bài toán thời gian thực, nơi quá trình phát hiện và theo dõi được tách biệt nhưng bổ trợ lẫn nhau. Luồng xử lý dữ liệu của hệ thống được mô tả chi tiết qua các giai đoạn sau:

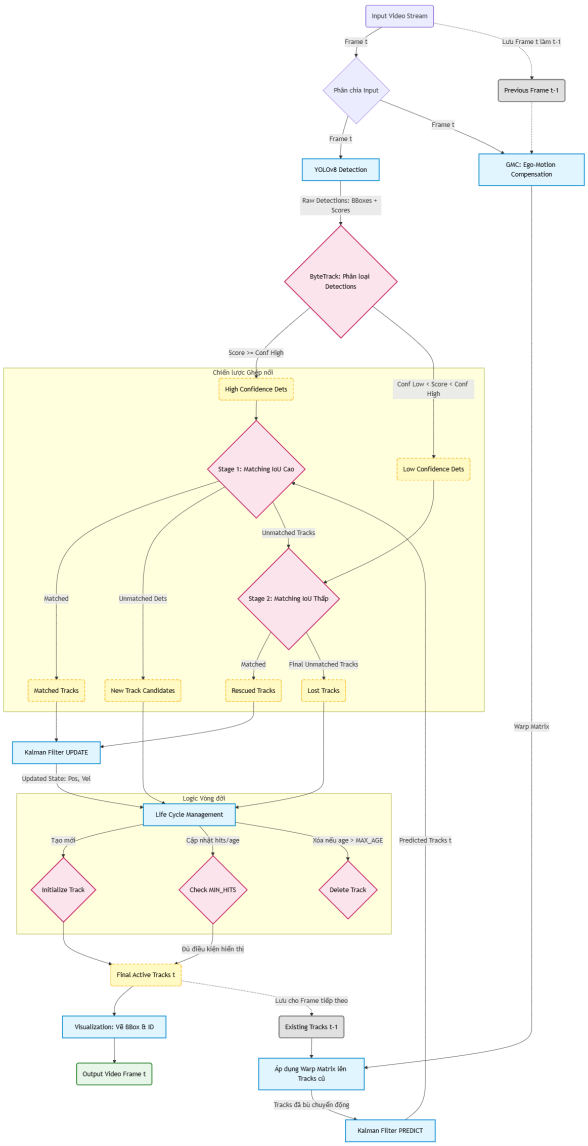
1. **Tiền xử lý Phát hiện:** Ảnh nhiệt đầu vào I_t tại thời điểm t được đưa qua mô hình phát hiện (đã tối ưu hóa Backbone như trình bày ở Mục 3.2) để trích xuất tập hợp các hộp bao $D_t = \{b_i, s_i\}$, với b_i là tọa độ và s_i là độ tin cậy.

2. **Ước lượng chuyển động nền (GMC):** Tính toán ma trận biến đổi giữa I_{t-1} và I_t để bù trừ sự dịch chuyển của camera UAV, đồng bộ hóa hệ tọa độ của các track cũ về thời điểm hiện tại.

3. **Dự đoán trạng thái (Kalman Prediction):** Sử dụng bộ lọc Kalman để dự đoán vị trí mới của các track dựa trên mô hình vận tốc không đổi.

4. **Liên kết dữ liệu (Data Association):** Sử dụng thuật toán ByteTrack để ghép nối các phát hiện D_t với các track dự đoán, ưu tiên các phát hiện có độ tin cậy cao nhưng không bỏ qua các phát hiện yếu.

5. **Cập nhật Quản lý vòng đời:** Cập nhật trạng thái vector Kalman cho các track được ghép nối, khởi tạo track mới hoặc xóa bỏ các track đã mất dấu.



Hình 4.1: Sơ đồ khối chi tiết luồng xử lý của hệ thống Tracking-by-Detection đề xuất.

4.2 Mô hình phát hiện đối tượng

Nghiên cứu đã đánh giá các mô hình lai, phân thành hai nhóm chính dựa trên mục tiêu tối ưu hóa (tốc độ hoặc độ chính xác):

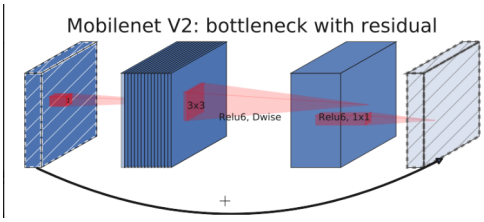
Bảng 4.1: Tổng hợp các mô hình Backbone được thử nghiệm

Nhóm	Backbone	Mô tả Kiến trúc	Mô hình
Lightweight/Mobile	**MobileNetV3**	Mạng siêu nhẹ, sử dụng *Mobile Inverted Bottleneck Convolution* (MBConv) và cơ chế Squeeze-and-Excitation (SE) Block.	MobileNetV3 + YOLOv8
	MobileNetV2	Sử dụng kiến trúc MBConv với Linear Bottlenecks để giảm thiểu chi phí tính toán.	MobileNetV2 + YOLOv8
	EfficientNetB3	Áp dụng Compound Scaling để mở rộng Depth, Width và Resolution một cách thống nhất.	EfficientNetB3 + YOLOv8
	EfficientNetB0	Mô hình cơ sở (baseline) của EfficientNet, tối ưu về tốc độ và tham số.	EfficientNetB0 + YOLOv8
Modern CNN	**ConvNeXt-T** (Tiny)	Thiết kế lại ResNet theo phong cách Vision Transformer (ViT), sử dụng Depthwise Convolution kích thước lớn (7x7).	ConvNeXt-T + YOLOv8
	ConvNeXt-S (Small)	Phiên bản mở rộng của ConvNeXt-T, cung cấp khả năng trích xuất đặc trưng mạnh mẽ hơn.	ConvNeXt-S + YOLOv8

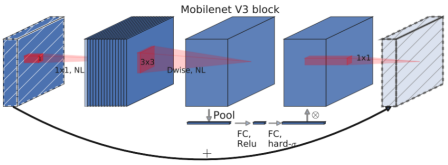
Để làm rõ hơn về lý do lựa chọn và cơ chế hoạt động, các kiến trúc Backbone cốt lõi được mô tả chi tiết qua hình minh họa:

4.2.1 MobileNet

Các phiên bản MobileNet được thiết kế tối ưu cho các thiết bị di động và nhúng bằng cách sử dụng khối cấu trúc MBConv (Mobile Inverted Bottleneck Convolution). Thay vì sử dụng tích chập tiêu chuẩn tốn kém, MBConv áp dụng kỹ thuật Depthwise Separable Convolution để tách rời quá trình xử lý không gian và kênh, giúp giảm đáng kể số lượng tham số và khối lượng tính toán. Đặc biệt, MobileNetV3 còn bổ sung các tối ưu hóa quan trọng như khối Squeeze-and-Excitation (SE) hoạt động như một cơ chế "chú ý", giúp mạng tự động tập trung vào các đặc trưng quan trọng và giảm nhiễu. Kết hợp cùng hàm kích hoạt nhẹ Hard-Swish và thuật toán tìm kiếm kiến trúc (NAS), mô hình đạt được sự cân bằng tốt nhất giữa độ chính xác và tốc độ xử lý thực tế.



(a) Cấu trúc MobileNet V2 Block



(b) Cấu trúc MobileNet V3 Block

Hình 4.2: So sánh cấu trúc MobileNet V2 và V3

4.2.2 EfficientNet

EfficientNet sử dụng chiến lược **Compound Scaling**, đồng thời mở rộng cả chiều sâu (Depth), chiều rộng (Width), và độ phân giải (Resolution) của mạng một cách thống nhất, nhằm tối đa hóa hiệu suất. . Việc thử nghiệm EfficientNetB0 (phiên bản baseline nhẹ nhất) và EfficientNetB3 (phiên bản cân bằng hơn) giúp đánh giá sự đánh đổi giữa kích thước và độ chính xác.

Bảng 4.2: Tóm tắt kiến trúc mạng

Stage	**Operator**	**Resolution**	**#Channels**	**#Layers**
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

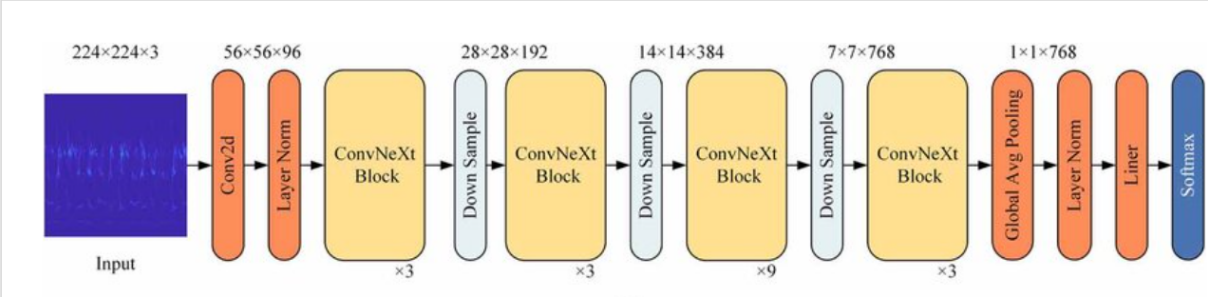
4.2.3 ConvNeXt

ConvNeXt được xem là mô hình CNN thế hệ mới, được "hiện đại hóa" bằng cách áp dụng các ý tưởng thiết kế chính từ **Vision Transformer (ViT)** . Nó thay thế các khối ResNet truyền thống bằng các khối đơn giản hơn, sử dụng Depthwise Convolution kích thước lớn (7x7), giúp cải thiện khả năng học ngữ cảnh mà không cần cơ chế Attention phức tạp.

Bảng 4.3: Kiến trúc ConvNeXt-Tiny với đầu vào 224×224×3

Giai đoạn	**Thành phần**	**Output size (H×W)**	**Channels**	**YOLOv8**
Input	Ảnh đầu vào	224 × 224	3	–
Stem	Conv 4×4, stride 4	56 × 56	96	–
Stage 1	3 × ConvNeXt Blocks	56 × 56	96	Không dùng
Stage 2	Downsampling + 3 × ConvNeXt Blocks	28 × 28	192	P3
Stage 3	Downsampling + 9 × ConvNeXt Blocks	14 × 14	384	P4
Stage 4	Downsampling + 3 × ConvNeXt Blocks	7 × 7	768	P5

Hình 4.3: Minh họa kiến trúc ConvNeXt



4.3 Bù trừ chuyển động toàn cục

Đặc thù của dữ liệu UAV là camera luôn chuyển động (pan, tilt, zoom) và rung lắc, dẫn đến việc vị trí pixel của một vật thể đứng yên cũng thay đổi liên tục. Nếu không xử lý, bộ lọc Kalman sẽ nhầm lẫn chuyển động của nền (ego-motion) là chuyển động của vật thể, gây ra sai số tích lũy lớn.

Chúng tôi áp dụng phương pháp bù trừ chuyển động dựa trên đặc trưng hình học (Image-based GMC) với các bước toán học như sau:

1. **Trích xuất đặc trưng:** Sử dụng thuật toán FAST để tìm các điểm đặc trưng nổi bật trên nền (background) của ảnh I_{t-1}.

2. **Theo dõi quang lưu:** Áp dụng phương pháp *Sparse Optical Flow* (Lucas-Kanade) để tìm vị trí các điểm này trên ảnh I_t.

3. **Ước lượng ma trận biến đổi:** Sử dụng giải thuật RANSAC để loại bỏ các điểm ngoại lai (outliers - thường là các điểm thuộc vật thể đang di chuyển) và tính toán ma trận biến đổi Affine $\mathbf{A}_t \in \mathbb{R}^{2\times 3}$:

$$\mathbf{A}_t = \begin{bmatrix} a_{11} & a_{12} & t_x \\ a_{21} & a_{22} & t_y \end{bmatrix} \tag{4.1}$$

   Trong đó a_{ij} đại diện cho phép xoay/co giãn và t_x, t_y là phép tịnh tiến.

4. **Đồng bộ hóa trạng thái:** Trước bước dự đoán của Kalman, vector trạng thái (vị trí trung tâm) của tất cả các track $\mathbf{x}_{t-1}$ sẽ được cập nhật về hệ tọa độ mới:

$$\begin{bmatrix} u'_{t-1} \\ v'_{t-1} \\ 1 \end{bmatrix} = \mathbf{A}_t \times \begin{bmatrix} u_{t-1} \\ v_{t-1} \\ 1 \end{bmatrix} \tag{4.2}$$

Việc áp dụng GMC giúp cải thiện đáng kể chỉ số IDF1 trong các kịch bản UAV bay nhanh hoặc xoay camera đột ngột.

4.4 Bộ lọc Kalman thích ứng

Để mô hình hóa chuyển động của đối tượng trong không gian 2D, chúng tôi sử dụng Bộ lọc Kalman với giả định vận tốc không đổi (Constant Velocity Model - CVM). Tuy nhiên, để đối phó với chất lượng phát hiện không ổn định của ảnh nhiệt, nghiên cứu tích hợp cơ chế **NSA (Noise Scale Adaptive)**.

4.4.1 Không gian trạng thái

Trạng thái của một track được định nghĩa bởi vector 8 chiều:

$$\mathbf{x} = [x_c, y_c, a, h, \dot{x}_c, \dot{y}_c, \dot{a}, \dot{h}]^T \tag{4.3}$$

Trong đó (x_c, y_c) là tọa độ tâm, a là tỷ lệ khung hình (w/h), h là chiều cao, và các biến có dấu chấm là đạo hàm theo thời gian (vận tốc).

4.4.2 Mô hình chuyển động và Đo lường

Quá trình dự đoán trạng thái từ bước $k-1$ sang k được thực hiện qua phương trình:

$$\mathbf{x}_{k|k-1} = \mathbf{F}_k \mathbf{x}_{k-1|k-1} + \mathbf{Q}_k \tag{4.4}$$

Với $\mathbf{F}$ là ma trận chuyển trạng thái (State Transition Matrix) mô tả mô hình vận tốc không đổi, và $\mathbf{Q}_k$ là ma trận hiệp phương sai nhiễu quá trình (Process Noise Covariance).

Quá trình cập nhật dựa trên đo lường $\mathbf{z}_k$ (hộp bao từ YOLO):

$$\mathbf{y}_k = \mathbf{z}_k - \mathbf{H}_k \mathbf{x}_{k|k-1} \tag{4.5}$$

Với $\mathbf{H}$ là ma trận đo lường, chỉ quan sát được 4 thành phần vị trí $[x_c, y_c, a, h]$.

4.4.3 Cơ chế thích ứng nhiễu

Điểm khác biệt trong hệ thống này là ma trận hiệp phương sai nhiễu đo lường $\mathbf{R}_k$ không cố định mà thay đổi dựa trên độ tin cậy (c_k) của hộp bao phát hiện và chiều cao đối tượng (h_k). Điều này xuất phát từ thực tế: hộp bao có điểm confidence thấp thường có vị trí kém chính xác hơn.

$$\mathbf{R}_k = (1 - c_k) \cdot \mathbf{R}_{std} \tag{4.6}$$

Cơ chế này giúp bộ lọc Kalman "tin tưởng" vào dự đoán của chính nó hơn khi gặp các phát hiện yếu (bị che khuất, mờ), giúp track mượt mà hơn (smoothness) trong điều kiện ảnh nhiệt

nhiễu.

4.5 Chiến lược liên kết dữ liệu ByteTrack

Khác với các phương pháp truyền thống (như DeepSORT) chỉ sử dụng các phát hiện có độ tin cậy cao (thường > 0.5), dẫn đến việc mất dấu các đối tượng nhỏ hoặc bị che khuất trong ảnh nhiệt, chúng tôi áp dụng chiến lược liên kết hai giai đoạn của ByteTrack.

Giải thuật được thực hiện như sau:

1. **Phân loại phát hiện:** Tập hợp các hộp bao D_t được chia thành hai tập con dựa trên ngưỡng τ_{high}:
   - D_{high}: Các phát hiện có điểm số $> \tau_{high}$ (Độ tin cậy cao).
   - D_{low}: Các phát hiện có điểm số trong khoảng $[\tau_{low}, \tau_{high}]$ (Độ tin cậy thấp, thường là đối tượng bị che khuất).
2. **Giai đoạn 1 (High Confidence Matching):** Thực hiện ghép nối các track hiện có (Tracks) với D_{high} sử dụng thuật toán Hungarian và độ đo khoảng cách IoU (Intersection over Union).

$$\text{IoU}(b_1, b_2) = \frac{\text{Area}(b_1 \cap b_2)}{\text{Area}(b_1 \cup b_2)} \tag{4.7}$$

   Các cặp ghép thành công sẽ được cập nhật trạng thái Kalman ngay lập tức.
3. **Giai đoạn 2 (Low Confidence Matching):** Các track chưa được ghép nối ở Giai đoạn 1 (T_{remain}) sẽ được thử ghép nối với tập D_{low}. *Lý do:* Trong ảnh nhiệt, khi đối tượng bị che khuất một phần, tín hiệu nhiệt giảm khiến confidence score giảm, nhưng vị trí không gian vẫn tương đồng. Việc tìm kiếm trong D_{low} giúp "cứu" các track này khỏi bị ngắt quãng.
4. **Xử lý ngoại lệ:** Các phát hiện trong D_{high} chưa được ghép sẽ khởi tạo track mới. Các track trong T_{remain} sau cả 2 giai đoạn sẽ bị đánh dấu là mất dấu (Lost).

4.6 Quản lý vòng đời

Để đảm bảo tính thời gian thực và giảm thiểu dương tính giả (False Positives), mỗi track được quản lý bởi một máy trạng thái (State Machine) với 3 trạng thái chính:

- **New (Mới):** Khi một phát hiện từ D_{high} không trùng với bất kỳ track nào, một Track mới được tạo với trạng thái *New*. Nếu Track này được ghép nối liên tiếp trong 3 khung hình (Init Frames), nó chuyển sang *Tracked*. Ngược lại, nếu mất dấu ngay lập tức, nó bị xóa (coi là nhiễu).
- **Tracked (Đang theo dõi):** Trạng thái ổn định của đối tượng. Tại đây, bộ đếm "thời gian mất dấu" (time_since_update) bằng 0.
- **Lost (Mất dấu):** Khi một Track không tìm thấy phát hiện tương ứng, nó chuyển sang trạng thái *Lost*. Tuy nhiên, hệ thống vẫn duy trì dự đoán vị trí của nó trong bộ nhớ thêm Max_Age khung hình (thường là 30 frames). Nếu đối tượng xuất hiện lại trong khoảng thời gian này, Track sẽ được khôi phục (Re-activated). Nếu vượt quá Max_Age, Track bị xóa vĩnh viễn (Deleted).

Cơ chế bộ đếm Max_Age là cực kỳ quan trọng đối với ảnh nhiệt UAV, nơi các vật thể thường xuyên bị che khuất bởi tán cây hoặc công trình trong thời gian ngắn.

Chương 5

Thực nghiệm và đánh giá kết quả

5.1 Kết quả thực nghiệm

Qua thực nghiệm trên một số mô hình, nhóm rút ra được bảng so sánh kết quả theo các chỉ số đánh giá như sau:

Bảng 5.1: So sánh hiệu suất các mô hình

Mô hình	**Backbone**	**Precision (P)**	**Recall (R)**	**F1-score**	**mAP@.50**	**mAP@.5:.95**
YOLOv8	CSPDarknet (YOLOv8-n)	0.89	0.74	0.808	0.808	0.521
MobileNetV3 + YOLOv8	MobileNetV3-L	0.812	0.737	0.773	0.787	0.521
MobileNetV2 + YOLOv8	MobileNetV2	0.789	0.737	0.762	0.772	0.503
EfficientNetB3 + YOLOv8	EfficientNetB3	0.82	0.737	0.776	0.787	0.53
EfficientNetB0 + YOLOv8	EfficientNetB0	0.914	0.765	0.833	0.811	0.521
ConvNeXt-T + YOLOv8	ConvNeXt-T	0.845	0.762	0.802	0.814	0.537
ConvNeXt-S + YOLOv8	ConvNeXt-S	0.87	0.785	0.825	0.861	0.572
YOLOv11	Custom/MBV3	0.793	0.561	0.657	0.655	0.404
MobileNetV3 + YOLOv10	MobileNetV3-L	0.7	0.565	0.625	0.601	0.338
MobileNetV2 + YOLOv10	MobileNetV2	0.75	0.636	0.688	0.699	0.425

Bảng 5.2: Hiệu suất tracking

Frames	**MOTA**	**MOTP**	**IDF1**	**MT**	**ML**	**ID Sw**
216	63.3%	0.308	80.9%	11	0	1

5.2 Đánh giá kết quả

- **mAP@.50 (Accuracy):** Mô hình **ConvNeXt-S + YOLOv8** đạt mức cao nhất (**0.861**), vượt qua cả YOLOv8 Default (**0.808**). Điều này chứng minh rằng Backbone thế hệ mới (ConvNeXt) là hiệu quả nhất để trích xuất đặc trưng trên bộ dữ liệu này. Trong nhóm Lightweight, EfficientNetB0 vẫn là tốt nhất và đạt chỉ số khá cao là(**0.811**).

- **mAP@.5:.95 (Độ Chính xác Định vị):** ConvNeXt-S cũng dẫn đầu ở chỉ số này (**0.572**), tiếp theo là ConvNeXt-T (**0.537**). Sự vượt trội này cho thấy các Backbone hiện đại mang lại đặc trưng chất lượng cao hơn, giúp YOLOv8 Head định vị hộp giới hạn chính xác hơn và ổn định hơn (mAP cao ở ngưỡng IoU nghiêm ngặt).

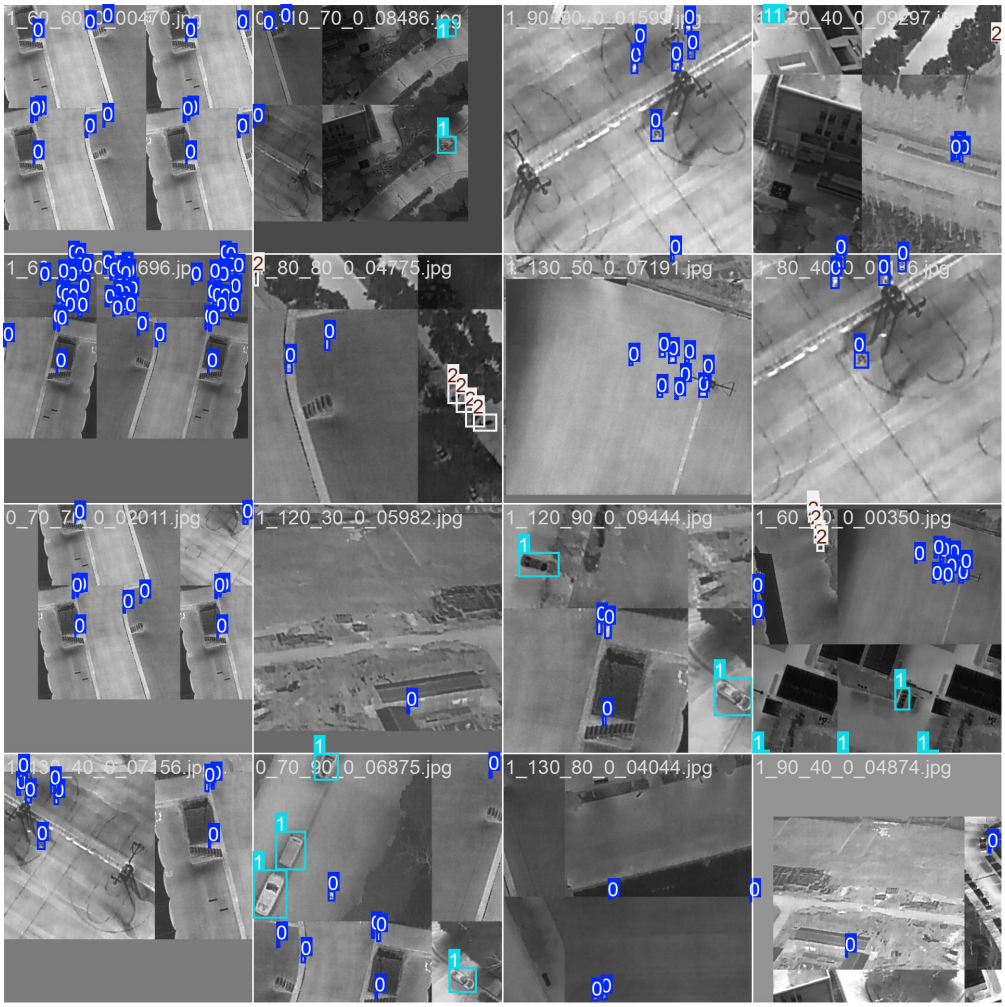
- **Precision (P):** Mô hình **EfficientNetB0** đã thể hiện hiệu suất vượt trội khi đạt Precision cao nhất (**0.914**), vượt qua mô hình gốc YOLOv8 Default (**0.890**) và các kiến trúc hiện đại khác. Kết quả này chứng minh rằng, mặc dù là Backbone nhẹ nhất, mô hình này hoạt động với mức **độ tin cậy** và **chọn lọc** rất cao, giảm thiểu tối đa các dự đoán sai (False Positives) trên bộ dữ liệu ảnh nhiệt.

- **Recall (R):** ConvNeXt-S (**0.785**) và ConvNeXt-T (**0.762**) tiếp tục thể hiện ưu thế về khả năng bao phủ đối tượng. Hiệu suất này khẳng định ConvNeXt là lựa chọn mạnh mẽ để trích xuất các đặc trưng phức tạp, khắc phục được điểm yếu Recall của YOLOv8 Default (**0.740**). Điều đáng nói là **EfficientNetB0** đạt kết quả ấn tượng là **0.765**, chỉ thấp hơn ConvNeXt-S một chút. Mức Recall này chứng tỏ EfficientNetB0, dù được tối ưu cho tốc độ, vẫn duy trì khả năng tìm kiếm đối tượng mạnh mẽ, đặc biệt là các đối tượng nhỏ trong ảnh UAV.

- **F1-score:** EfficientNetB0 đã đạt F1-score cao nhất (**0.833**), vượt qua ConvNeXt-S (**0.825**) và YOLOv8 Default (**0.808**). F1-score là thước đo tổng hợp thể hiện sự cân bằng giữa P và R. Việc EfficientNetB0 dẫn đầu chứng minh rằng sự kết hợp giữa Precision vượt trội và Recall mạnh mẽ đã giúp mô hình nhẹ này trở thành mô hình **cân bằng nhất và toàn diện nhất** về mặt thống kê trong tất cả các mô hình được thử nghiệm.

Dựa trên kết quả thực nghiệm mới nhất, mô hình ConvNeXt-S + YOLOv8 đạt hiệu suất cao

nhất về khả năng bao phủ đối tượng (Recall cao nhất), trong khi mô hình EfficientNetB0 + YOLOv8 lại đạt hiệu suất cao nhất về độ chính xác (Precision và F1-score cao nhất). Khi đánh giá sự cân bằng tối ưu giữa tốc độ và độ chính xác (Latency/Accuracy Trade-off), mô hình EfficientNetB0 + YOLOv8 thể hiện tính hiệu quả vượt trội. Mô hình lai này sử dụng Backbone nhẹ nhất để giảm thiểu độ trễ (Latency), nhưng lại đạt F1-score cao nhất. Do đó, EfficientNetB0 + YOLOv8 là mô hình tối ưu và cân bằng nhất trong số các mô hình đã thử nghiệm.



Hình 5.1: Efficientnet B0 train batch 0

Đánh giá Hiệu suất Tracking (Multi-Object Tracking Results):

Kết quả thực nghiệm trên tập dữ liệu video kiểm thử (gồm 216 khung hình) cho thấy hệ thống đề xuất hoạt động ổn định với các chỉ số ấn tượng:

- **IDF1 (ID F1 Score):** Hệ thống đạt mức **80.9**%. Đây là chỉ số quan trọng nhất phản

ánh khả năng duy trì định danh (Identity) của đối tượng theo thời gian. Mức điểm cao này chứng minh hiệu quả của chiến lược liên kết **ByteTrack** kết hợp với bộ lọc **NSA-Kalman**, giúp hệ thống không bị "nhầm lẫn" đối tượng ngay cả khi chúng di chuyển gần nhau hoặc bị che khuất một phần.

- **MOTA (Multiple Object Tracking Accuracy):** Đạt **63.3**%. Chỉ số này tổng hợp các lỗi sai sót (False Negatives, False Positives, ID Switches). Mặc dù MOTA bị ảnh hưởng mạnh bởi chất lượng Detection (đặc biệt là các lỗi phát hiện sai do nhiễu ảnh nhiệt), nhưng mức trên 60% vẫn được coi là kết quả khả quan cho bài toán giám sát UAV phức tạp.

- **ID Switches (ID Sw):** Chỉ số chuyển đổi ID cực thấp, chỉ có **1** lần xảy ra sự cố đổi nhầm ID trong suốt quá trình theo dõi. Điều này khẳng định tính bền vững (Robustness) của thuật toán quản lý vòng đời quỹ đạo, đặc biệt là cơ chế **Global Motion Compensation (GMC)** đã loại bỏ hiệu quả các nhiễu động do rung lắc camera.

- **MT (Mostly Tracked) và ML (Mostly Lost):**
  - Số lượng quỹ đạo được theo dõi trọn vẹn (MT) là **11**, chiếm tỷ lệ áp đảo.
  - Số lượng quỹ đạo bị mất hoàn toàn (ML) là **0**.

  Kết quả này cho thấy hệ thống có khả năng "bám đuổi" mục tiêu rất tốt, không để lọt bất kỳ đối tượng quan trọng nào.

Chương 6

Kết luận và hướng phát triển

Hệ thống phát hiện mục tiêu trên ảnh nhiệt tầm cao từ UAV sử dụng YOLOv8 kết hợp với custom backbone đã cho thấy hiệu quả vượt trội trong việc nhận dạng các đối tượng nhỏ và ít chi tiết trên nền nhiệt độ phức tạp. Việc thay thế backbone mặc định bằng các kiến trúc tối ưu hơn như EfficientNet hoặc MobileNet giúp mô hình trích xuất đặc trưng rõ nét hơn, cải thiện mAP và đồng thời duy trì tốc độ xử lý thời gian thực phù hợp cho triển khai trên nền tảng UAV. Mô hình hoạt động ổn định ở nhiều độ cao, trong nhiều điều kiện quan sát khác nhau, chứng minh khả năng tổng quát hóa tốt. Trong tương lai, hệ thống có thể được mở rộng theo các hướng như tích hợp bộ theo dõi mục tiêu (tracking) để tăng độ ổn định, kết hợp đa nguồn dữ liệu như RGB hoặc bản đồ độ cao để giảm nhiễu, áp dụng chiến lược tăng cường dữ liệu chuyên biệt cho ảnh nhiệt, hoặc tối ưu hóa mô hình bằng TensorRT để triển khai trên các thiết bị nhúng. Những hướng phát triển này sẽ giúp nâng cao hơn nữa độ chính xác, độ tin cậy và tính ứng dụng thực tế của hệ thống trong các nhiệm vụ giám sát và nhận dạng mục tiêu từ UAV.

Tài liệu tham khảo

[1] MobileNet Architecture Series (Viblo). *CNN Architecture Series 1 – MobileNets: Mô hình gọn nhẹ cho Mobile Applications*. https://viblo.asia/p/cnn-architecture-series-1-mobilenets-mo-hinh-gon-nhe-cho-mobile-applications-1VgZv

[2] ConvNeXt Architecture (GeeksforGeeks). https://www.geeksforgeeks.org/computer-vision/convnext/

[3] EfficientNet Architecture (GeeksforGeeks). https://www.geeksforgeeks.org/computer-vision/efficientnet-architecture/

[4] HIT-UAV Dataset (Kaggle). https://www.kaggle.com/datasets/pandrii000/hituav-a-highaltitude-infrared-thermal-dataset