BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT

THÀNH PHỐ HỒ CHÍ MINH

NGÔ ĐỨC ĐẠT

NHẬN DẠNG VÀ PHÂN LOẠI TÀU TRONG CẢNH GIỚI BỜ BIỂN SỬ DỤNG TRÍ TUỆ NHÂN TẠO

LUẬN ÁN TIẾN SĨ

NGÀNH: KỸ THUẬT ĐIỆN TỬ

Tp. Hồ Chí Minh, tháng 4 năm 2026

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT

THÀNH PHỐ HỒ CHÍ MINH

NGÔ ĐỨC ĐẠT

NHẬN DẠNG VÀ PHÂN LOẠI TÀU TRONG CẢNH GIỚI BỜ BIỂN SỬ DỤNG TRÍ TUỆ NHÂN TẠO

LUẬN ÁN TIẾN SĨ

NGÀNH: KỸ THUẬT ĐIỆN TỬ

Người hướng dẫn khoa học 1: PGS.TS. LÊ MỸ HÀ

Người hướng dẫn khoa học 2: PGS.TS. NGUYỄN MẠNH HÙNG

Phản biện 1:

Phản biện 2:

Tp. Hồ Chí Minh, tháng 4 năm 2026

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH

Số: 256/QĐ-ĐHSPKT

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

Tp. Hồ Chí Minh, ngày 02 tháng 02 năm 2021

QUYẾT ĐỊNH

Về việc giao đề tài luận án và người hướng dẫn NCS khóa 2020 - 2023

HIỆU TRƯỞNG TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

Căn cứ Luật Giáo dục đại học ngày 18/6/2012 và Luật sửa đổi, bổ sung một số điều của Luật Giáo dục đại học ngày 19/11/2018;

Căn cứ Nghị định 99/2019/NĐ-CP ngày 30/12/2019 của Chính phủ Quy định chi tiết và hướng dẫn thi hành một số điều của Luật sửa đổi, bổ sung một số điều của Luật giáo dục đại học;

Căn cứ Quyết định số 937/QĐ-TTg ngày 30/6/2017 của Thủ tướng Chính phủ về việc phê duyệt đề án thí điểm đổi mới cơ chế hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Nghị quyết số 11/NQ-HĐT ngày 08/01/2021 của Hội đồng trường ban hành Quy chế tổ chức và hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. HCM;

Căn cứ Thông tư số 08/2017/TT-BGDĐT ngày 04/4/2017 của Bộ Giáo dục và Đào tạo về việc Ban hành Qui chế tuyển sinh và đào tạo trình độ tiến sĩ;

Theo nhu cầu công tác và khả năng cán bộ;

Theo đề nghị của Trưởng khoa/Viện quản ngành và Trưởng phòng Đào tạo,

QUYẾT ĐỊNH:

Điều 1. Giao đề tài luận án tiến sĩ và người hướng dẫn cho:

Nghiên cứu sinh : ***Ngô Đức Đạt***

Ngành : ***Kỹ thuật điện tử*** Khoá: ***2020 – 2023***

Tên luận án : ***Nhận dạng và phân loại tàu biển từ tín hiệu radar***

Người HD thứ nhất (HD chính) : ***PGS.TS. Lê Mỹ Hà***

Người HD thứ hai : ***TS. Nguyễn Mạnh Hùng***

Thời gian thực hiện : ***01/6/2020 đến 31/5/2023***

Điều 2. Giao cho Phòng Đào tạo quản lý, thực hiện theo đúng Qui chế đào tạo trình độ tiến sĩ của Bộ Giáo dục & Đào tạo và Nhà trường đã ban hành.

Điều 3. Trưởng các đơn vị, phòng Đào tạo, các Khoa/Viện quản ngành tiến sĩ và các Ông (Bà) có tên tại Điều 1 chịu trách nhiệm thi hành quyết định này.

Quyết định có hiệu lực kể từ ngày ký./.

Nơi nhận:
- BGH (để biết);
- Như điều 2, 3;
- Lưu: VT, ĐT (4b).

HIỆU TRƯỞNG

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

PGS.TS. Đỗ Văn Dũng

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT
THÀNH PHỐ HỒ CHÍ MINH

Số: 1130 /QĐ-ĐHSPKT

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

Tp. Hồ Chí Minh, ngày 26 tháng 4 năm 2023

QUYẾT ĐỊNH
Về việc đổi tên luận án tiến sĩ

HIỆU TRƯỞNG TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

Căn cứ Luật Giáo dục đại học ngày 18/6/2012 và Luật sửa đổi, bổ sung một số điều của Luật Giáo dục đại học ngày 19/11/2018;

Căn cứ Nghị định 99/2019/NĐ-CP ngày 30/12/2019 của Chính phủ Quy định chi tiết và hướng dẫn thi hành một số điều của Luật sửa đổi, bổ sung một số điều của Luật giáo dục đại học;

Căn cứ Quyết định số 937/QĐ-TTg ngày 30/6/2017 của Thủ tướng Chính phủ về việc phê duyệt đề án thí điểm đổi mới cơ chế hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Nghị quyết số 11/NQ-HĐT ngày 08/01/2021 của Hội đồng trường ban hành Quy chế tổ chức và hoạt động của Trường Đại học Sư phạm Kỹ thuật TP. HCM;

Căn cứ Nghị quyết số 99/NQ-HĐT ngày 21/10/2022 của Hội đồng trường về công tác cán bộ lãnh đạo Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Nghị quyết số 118/NQ-HĐT ngày 27/01/2023 của Hội đồng trường sửa đổi, bổ sung Quy chế tổ chức và hoạt động của trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh;

Căn cứ Thông tư số 08/2017/TT-BGDĐT ngày 04 tháng 4 năm 2017 của Bộ Giáo dục và Đào tạo về việc Ban hành Qui chế tuyển sinh và đào tạo trình độ tiến sĩ;

Căn cứ Quyết định số 1557/QĐ-ĐHSPKT ngày 30/8/2017 của Trường Đại học Sư phạm Kỹ thuật TP. Hồ Chí Minh về việc ban hành Quy định đào tạo tiến sĩ của trường đại học Sư phạm Kỹ thuật Tp.HCM;

Theo đề nghị của Nghiên cứu sinh, Khoa quản ngành và Trưởng phòng Đào tạo.

QUYẾT ĐỊNH:

Điều 1. Đổi tên luận án tiến sĩ cho:

Nghiên cứu sinh : ***Ngô Đức Đạt***
Ngành : Kỹ thuật điện tử Khoá: ***2020 – 2023***
Tên luận án mới : ***Nhận dạng và phân loại tàu trong cảnh giới bờ biển sử dụng trí tuệ nhân tạo***
Người HD thứ nhất (HD chính): ***PGS.TS. Lê Mỹ Hà***
Người HD thứ hai : ***TS. Nguyễn Mạnh Hùng***
Thời gian thực hiện : ***01/6/2020 đến 31/5/2023***

Điều 2. Giao cho Phòng Đào tạo quản lý, thực hiện theo đúng Qui chế đào tạo trình độ tiến sĩ của Bộ Giáo dục & Đào tạo và Nhà trường đã ban hành.

Điều 3. Trưởng các đơn vị: phòng Đào tạo, khoa/viện quản ngành, phòng KHTC và các Ông (Bà) có tên ở Điều 1 chịu trách nhiệm thi hành quyết định này./.

Nơi nhận:
- BGH (để chỉ đạo);
- Như điều 3;
- Lưu: VT, ĐT (3b).

KT. HIỆU TRƯỞNG
PHÓ HIỆU TRƯỞNG

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

PGS.TS. Lê Hiếu Giang

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của tôi.

Các số liệu, kết quả nêu trong Luận án là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác

Tp. Hồ Chí Minh, ngày 9 tháng 4 năm 2026

Ngô Đức Đạt

LỜI CẢM ƠN

Trước tiên, tôi xin được bày tỏ lòng biết ơn, gửi lời cảm ơn chân thành và sâu sắc đến thầy **PGS. TS. Lê Mỹ Hà**, thầy **PGS.TS Nguyễn Mạnh Hùng**, hai người Thầy luôn rất nhiệt tình và tận tâm hướng dẫn tôi trong thời gian thực hiện chuyên đề. Hơn nữa, trong suốt quá trình thực hiện từ lúc lập đề cương cho đến khi thực hiện luận án, Thầy luôn có những góp ý và định hướng và tận tình giúp tôi đạt được những kết quả tốt nhất.

Tôi cũng đồng gửi lời cảm ơn đến quý Thầy Cô ở Trường Đại Học Sư Phạm Kỹ Thuật Tp. HCM nói chung cũng như các Thầy Cô ở khoa Điện – Điện tử nói riêng đã truyền đạt cho tôi những nền tảng kiến thức quý báu, giúp tôi có thể hoàn thành được chuyên đề.

Tôi cũng xin gửi lời cảm ơn sâu sắc đến Viện sau đại học, Viện luôn nhiệt tình hường dẫn thực hiện các thủ tục, hồ sơ để tôi hoàn thiện luận án kịp thời và đúng mẫu biểu.

Cuối cùng, tôi cũng xin được gửi lời biết ơn sâu sắc đến gia đình tôi, những người đã luôn là chỗ dựa tinh thần, là nguồn động viên vô cùng to lớn trong những lúc khó khăn, giúp tôi có thể an tâm thực hiện công việc học tập và nghiêm cứu của mình trong suốt thời gian thực hiện chuyên đề.

Xin chân thành cảm ơn!

Tp. Hồ Chí Minh, tháng 4 năm 2026

NCS

NGÔ ĐỨC ĐẠT

MỤC LỤC

MỤC LỤC i

DANH SÁCH BẢNG iv

DANH SÁCH HÌNH VẼ vi

DANH MỤC TỪ VIẾT TẮT ix

TÓM TẮT LUẬN ÁN x

ABSTRACT 1

1 TỔNG QUAN 6

1.1 Giới thiệu 6

1.2 Các công bố quốc tế về giám sát an ninh bờ biển 8

1.3 Các công bố trong nước về giám sát an ninh bờ biển 10

1.4 Kết luận của chương 12

2 CƠ SỞ LÝ THUYẾT 15

2.1 Cơ sở lý thuyết về phát hiện đối tượng bằng cảm biến ảnh 15

2.1.1 Hệ thống camera giám sát bờ biển 15

2.1.2 Phát hiện đối tượng trong ảnh 17

2.1.3 Phát hiện đối tượng bằng đặc trưng thủ công 18

2.1.4 Phát hiện đối tượng dựa trên mô hình học sâu 19

2.1.5 Tổng quan các phương pháp phát hiện đối tượng 21

2.1.6 Kiến trúc YOLOX và cơ chế không điểm neo (Anchor-free) 24

2.1.7 mô hình phát hiện đối tượng dựa trên Transformer 25

2.1.8 Các chỉ số đánh giá cho bài toán phát hiện đối tượng đối tượng 28
2.2 Phân cụm bằng kỹ thuật học sâu 32
2.2.1 Tổng quan về phân cụm 32
2.2.2 Phân cụm dữ liệu truyền thống 33
2.2.3 Phân cụm dữ liệu dựa trên học sâu 34
2.2.4 Các chỉ số đánh giá 35
2.3 Học đặc trưng không giám sát 39
2.3.1 Mạng nơ-ron tự mã hóa (Autoencoder - AE) 39
2.3.2 Mạng nơ-ron tự mã hóa biến phân (VAE) 40
2.4 Cơ sở lý thuyết về phân loại tín hiệu 41
2.4.1 Mạng Nơ-ron nhân tạo cho bài toán phân loại 41
2.4.2 Các chỉ số đánh giá cho bài toán phân loại đối tượng 43

3 PHÁT HIỆN TÀU BIỂN TỪ DỮ LIỆU ẢNH 46
3.1 Phát hiện tàu biển bằng mô hình Transformer 46
3.2 Các phương pháp phát hiện tàu biển 47
3.3 Phương pháp phát hiện tàu dựa trên mạng YOLOX kết hợp VIB 48
3.3.1 Lựa chọn đặc trưng bằng VIB 55
3.3.2 Tổng quan Phương pháp đề xuất 58
3.3.3 Khối trích xuất đặc trưng và khối kết nối 60
3.4 Bộ dữ liệu và ngữ cảnh thử nghiệm 60
3.5 Kết quả thực nghiệm với mạng YOLOX 62
3.5.1 Chọn siêu tham số α_{KL} 62
3.5.2 Ảnh hưởng của việc lựa chọn các bộ rút trích đặc trưng 63
3.5.3 Ảnh hưởng của vị trí của khối VIB trong mô hình phát hiện đối tượng 66
3.5.4 So sánh với các phương án tốt nhất trong phát hiện tàu biển 68
3.5.5 Thí nghiệm trên các tập dữ liệu nhỏ 69
3.6 Kết quả thí nghiệm với DETR 70
3.6.1 Lựa chọn siêu tham số 71
3.6.2 So sánh với các phương pháp tiên tiến nhất (SoTA) 73
3.6.3 Nghiên cứu cắt giảm các hàm mất mát 76
3.6.4 Phân tích đặc trưng 77

3.7 Kết luận của chương 78

4 PHÂN TÍCH DỮ LIỆU RADAR BẰNG KỸ THUẬT HỌC SÂU 82

4.1 Phân tích đặc trưng tín hiệu Radar và phương pháp phân cụm 82

4.1.1 Thách thức trong phân tích dữ liệu radar thực tế 83

4.1.2 Đề xuất phương pháp trích xuất đặc trưng kết hợp 83

4.1.3 Mục tiêu nghiên cứu của chương 84

4.2 Phân đoạn và trích xuất đặc trưng 85

4.2.1 Tổng quan về hệ thống 85

4.2.2 Các thuộc tính được đề xuất bởi chuyên gia 86

4.2.3 Rút trích đặc trưng 88

4.2.4 mô hình rút trích đặc trưng và hàm mục tiêu 89

4.2.5 Bộ dữ liệu và quá trình thu thập dữ liệu 90

4.2.6 Kết quả thí nghiệm với các phương pháp truyền thống 92

4.2.7 So sánh với các phương pháp học sâu tiên tiến 93

4.3 mô hình và đánh giá 95

4.3.1 Rút trích đặc trưng 96

4.3.2 Các mô hình phân loại 97

4.3.3 Bộ phân loại dựa trên học sâu 97

4.3.4 Tổng quan hệ thống 98

4.3.5 Phân cụm đối tượng Kmeans 98

4.4 Một số kết quả thí nghiệm 100

4.4.1 So sánh các đặc trưng trong miền tần số và trong miền thời gian 100

4.4.2 Ảnh hưởng của thuật toán phân nhóm lên bài toán phân loại 103

4.5 Kết luận của chương 105

5 KẾT LUẬN VÀ KIẾN NGHỊ 107

5.1 Kết luận 107

5.2 Kiến nghị 110

TÀI LIỆU THAM KHẢO 111

DANH MỤC CÔNG TRÌNH CỦA NGHIÊN CỨU SINH LIÊN QUAN ĐẾN ĐỀ TÀI 118

DANH SÁCH BẢNG

2.1 So sánh các chỉ số đánh giá phân cụm 39

3.1 Ký hiệu toán học 58

3.2 Chi tiết mạng. Ở đây, $Encoder_{\mu}$ là bộ mã hóa để trích xuất trị trung bình μ, $Encoder_{\sigma}$ là bộ mã hóa để trích xuất độ lệch chuẩn σ, $i$là chỉ số của cấp độ quy mô, và C^i là số lượng kênh trong đầu vào của cấp độ i^{th}. 60

3.3 So sánh mAP có và không có VIB trên các bộ rút trích đặc trưng khác nhau. 64

3.4 So sánh mAP và tốc độ suy luận (FPS) có và không có VIB trên các bộ rút trích đặc trưng khác nhau. 64

3.5 Đánh giá hiệu quả dựa trên vị trí chèn mô-dun VIB. 67

3.6 So sánh hiệu suất của các phương pháp khác nhau. Các kết quả tốt nhất được **in đậm**. 69

3.7 Hiệu suất trên các tập dữ liệu nhỏ. S_1 nghĩa là 30% mẫu huấn luyện, S_2 nghĩa là 70% mẫu huấn luyện, S_3 nghĩa là 100% mẫu huấn luyện từ D_2^{Train}. Các kết quả tốt nhất được **in đậm**. 70

3.8 Ảnh hưởng của số lượng truy vấn đối tượng(*queries*) tới kết quả phát hiện đối tượng. 72

3.9 So sánh hiệu năng của mô hình Deformable DETR với learning rates khác nhau. Các kết quả tốt nhất được in đậm. 72

3.10 So sánh hiệu suất của các phương pháp khác nhau. Kết quả tốt nhất được in đậm. . . 74

3.11 Hiệu suất trên các tập dữ liệu nhỏ. S_1 nghĩa là 30% mẫu huấn luyện, S_2 nghĩa là 70% mẫu huấn luyện, S_3 nghĩa là 100% mẫu huấn luyện từ D_2^{Train}. 76

3.12 So sánh độ phức tạp giữa VIB-detector và DETR-detector. 76

3.13 So sánh mAP của Deformable DETR khi giảm các hàm mất mát huấn luyện. 77

4.1 Kết quả phân cụm dựa vào Kmeans và các đặc trưng. 92

4.2 Kết quả phân cụm bằng kỹ thuật học sâu 93

4.3 Kết quả phân loại với các đặc tính khác nhau của các loại tàu. 101
4.4 Tóm tắt các cấu Hình mạng Nơ-ron và thông số huấn luyện 102
4.5 Kết quả phân loại bằng các đặc trưng khác nhau với tỷ lệ huấn luyện/kiểm tra là 2. . . 102

DANH SÁCH HÌNH VẼ

2.1 Cấu tạo của thiết bị quang điện tử. (1) Khối Camera nhiệt; (2) Khối Camera màu; (3) Khối chân đế và khối quay; (4) Khối nguồn 16
2.2 Mối quan hệ giữa các bài toán nhận dạng đối tượng trong thị giác máy tính 18
2.3 Bài toán phát hiện nhiều đối tượng trong ảnh 19
2.4 Phát hiện đối tượng dựa trên vùng đề xuất và phát hiện đối tượng dựa trên hồi quy/phân loại 20
2.5 Bài toán phát hiện đơn đối tượng trong ảnh 20
2.6 Khung bao để phát hiện khung bao cụ thể 21
2.7 Kiến trúc DETR 25
2.8 Intersection over Union 29
2.9 Ví dụ minh họa đường cong Precision–Recall [10] 30
2.10 Sơ đồ nguyên lý hoạt động của mạng AE. 40
2.11 Sơ đồ nguyên lý hoạt động của mạng VAE. 41
2.12 Sơ đồ khối một mạng Neuron 42
2.13 Ma trận nhầm lẫn. 44
2.14 Chỉ số đánh giá phân loại ROC-AUC 45

3.1 Hình minh họa Đường dẫn từ dưới lên 49
3.2 Cấu trúc lan truyền từ dưới lên và từ trên xuống của FPN 49
3.3 Minh họa đường đi theo từ dưới lên và từ trên xuống 50
3.4 Cấu trúc của Dilated Encoder [1] 51
3.5 Cấu trúc của YOLOF [1] 53
3.6 Minh hoạ đầu gộp và đầu tách [2] 54
3.7 Sơ đồ khối tổng thể hệ thống nhận dạng và phân loại tàu biển 59

3.8 kiến trúc mạng: (a) tổng quan phương pháp phát hiện đối tượng dựa trên VIB; (b) nhánh phân loại dựa trên VIB được đề xuất. 59
3.9 Cấu trúc Darknet 61
3.10 Cấu trúc PAFPN 62
3.11 Giá trị mAP ứng với các tham số α_{VIB} khác nhau. Trục x có nghĩa mô tả tham số α_{KL}, và trục y mô tả độ chính xác trung bình trung bình trên tất cả các lớp. 63
3.12 Bản đồ đặc trưng trên lớp trung gian của bộ phân loại 65
3.13 Bản đồ đặc trưng trên tầng đầu ra của bộ phân loại 65
3.14 VIB ở phần giữa của đầu phân loại 66
3.15 VIB ở phần bắt đầu của đầu tách 66
3.16 Giá trị hàm mục tiêu trong một quá trình huấn luyện. Dòng in đậm là các giá trị được làm trơn qua các lần lặp. Đường nhạt hơn là giá trị thực tế trong một lần lặp. 67
3.17 Hiệu suất khi số lượng mẫu huấn luyện bị giới hạn. "Small" nghĩa là 30% mẫu huấn luyện, "Medium" nghĩa là 70% mẫu huấn luyện, và "Large" nghĩa là 100% mẫu huấn luyện từ D_2^{Train}. 75
3.18 Bản đồ đặc trưng tại đầu phân loại, phần cổ và phần xương sống. (Văn bản trên ảnh gốc đã được loại bỏ.) 78

4.1 Ba nhóm dạng sóng thu được từ thực địa cho thấy sự chồng lấn về biên độ. 83
4.2 Phân tách đặc trưng phần đỉnh và phần đáy của xung phản hồi dựa trên tri thức chuyên gia. 84
4.3 Tổng quan về các phương pháp được đề nghị. 86
4.4 Dạng sóng phản hồi đặc trưng của tàu cá (biên độ nhỏ, thường xuất hiện nhiều đỉnh sóng do đi theo cụm). 87
4.5 Dạng sóng phản hồi của tàu quân sự theo các khoảng cách khác nhau. 87
4.6 Dạng sóng phản hồi của tàu vận tải theo hướng di chuyển và khoảng cách. 87
4.7 lưu đồ minh hoạ cho phương pháp xử lý tín hiêu radar. 88
4.8 Tiền xử lý tín hiệu radar để rút trích đặc trưng. 88
4.9 Màn Hình điều khiển khi xuất hiện mục tiêu 90
4.10 Dạng sóng của tín hiệu phản xạ và tệp dữ liệu. 91
4.11 Ảnh hưởng của hệ số α tới chỉ số mutual information. 95
4.12 Sơ đồ xử lý tín hiệu phản xạ từ mục tiêu 96

4.13 Tổng quan hệ thống 98
4.14 Độ tăng cường của ACC và F1-Score khi tỷ lệ huấn luyện/testing là 66%/33% 104
4.15 Độ tăng cường của ACC và F1-Score khi tỷ lệ huấn luyện/testing là 50%/50% 104

DANH MỤC TỪ VIẾT TẮT

AUC	Area Under the Curve
TPR	True Positive Rate
ROC	Receiver Operating Characteristic
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
CNN	Convolutional Neural Network
NN	Neuron Network
RCS	Radar Cross Section
VIB	Variational Information Bottleneck
FPN	Feature Pyramid Network
FFT	Fast Fourier Transform
DCT	Discrete Cosine Transform
PAFPN	Path Aggregation Feature Pyramid Network
KL	Kullback–Leibler
SGD	Stochastic Gradient Descent

TÓM TẮT

Nội dung chính của luận án là đề xuất mô hình học sâu cải tiến cho việc phát hiện đối tượng tàu biển bằng cách sử dụng các cảm biến hình ảnh và cảm biến radar xung. Các mô hình sử dụng kết hợp cả các dữ liệu được công bố rộng rãi trong việc phát hiện tàu biển bằng hình ảnh và đồng thời xây dựng một bộ dữ liệu sóng radar xung thực tế thu được từ các trạm cảnh giới biển của Việt Nam.

Có hai mô hình phát hiện tàu từ dữ liệu hình ảnh. Mô hình đầu tiên dựa trên các kiến trúc mạng nơ ron nhân chập và tuỳ biến thêm các khối lựa chọn đặc trưng. Các khối lựa chọn đặc trưng này rút trích một lượng thông tin ít nhất có thể để phát hiện tàu biển và loại bỏ những thông tin thừa. Từ đó đạt được các đặc trưng có độ chính xác cao hơn. Kết quả thí nghiệm với bộ dữ liệu kích thước lớn và được công bố rộng rãi cho thấy mô hình có thể đạt được độ chính xác cao hơn ngay cả khi sử dụng ít dữ liệu huấn luyện hơn so với các phương pháp tiên tiến khác. Không những thế, sự cải tiến sẽ càng rõ ràng hơn nếu số lượng các mẫu huấn luyện giảm dần. Các phân tích từ bản đồ đặc trưng cũng cho thấy đáp ứng với bản đồ đặc trưng tập trung tốt hơn so với các phương pháp truyền thống. Mô hình thứ hai là mô hình phát hiện đối tượng dựa trên transformer. Các mô hình dựa trên transformer có cơ chế tập trung vào các điểm quan trọng. Cơ chế này cũng hoạt động tương tự như lựa chọn đặc trưng. Nhờ đó, mô hình đạt độ chính xác cao hơn các phương pháp dựa trên mạng CNN. Sự cải thiện này đặc biệt rõ ràng hơn khi mô hình được huấn luyện với lượng nhỏ dữ liệu. Bản đồ tập trung vào các điểm ảnh thậm chí còn tốt hơn mô hình lựa chọn đặc trưng trước đó trên các khối CNN.

Đối với tín hiệu radar, luận án thu thập dữ liệu từ một trạm giám sát dọc bờ biển Việt Nam. Các kiến thức chuyên gia được sử dụng để hướng dẫn mô hình học sâu rút trích các đặc trưng dựa theo sự hướng dẫn của chuyên gia. Sau đó, một phương pháp phân loại tàu từ tín hiệu radar dựa trên các kết quả phân cụm sẽ được trình bày. So với các phương pháp truyền thống, phương pháp này thể hiện ba đóng góp chính. Đóng góp đầu tiên là bộ dữ liệu thực tế được thu từ một trạm radar giám sát tại Việt Nam. Khác với các nghiên cứu trước đây nơi dữ liệu chủ yếu được mô phỏng, dữ liệu này được thu thập thực tế và có giá trị sử dụng cao. Các đặc trưng từ kinh nghiệm của các chuyên gia giám sát biển cũng được sử dụng để phân loại thuyền. Các đặc trưng này được đánh giá qua một thuật toán phân nhóm. Cuối cùng, thuật toán phân loại thuyền dựa trên tín hiệu radar được giới thiệu dựa trên kết quả phân nhóm và đạt được độ chính xác cao.

ABSTRACT

The thesis proposes deep learning models for maritime vessel detection based on camera and pulse radar sensors. The proposed models are evaluated using both published image datasets and a real-world pulse radar dataset that was collected from coastal surveillance stations in Vietnam.

In camera based detection, two deep-learning-based models have been investigated for ship detection. The first model is built upon convolutional neural network (CNN) architectures. The baseline model (YoloX) is customized by incorporating feature selection modules. These feature selection modules are designed to extract compressed features that are suitable for vessel detection while suppressing redundant and irrelevant information. Therefore, it could learn discriminative feature representations for a task. Experimental results on large-scale datasets demonstrate that the proposed model achieves higher detection accuracy with significantly fewer training samples. Moreover, the performance gain becomes more pronounced as the number of training samples decreases. Feature-maps indicate that the proposed model exhibits more focused and discriminative activation patterns than conventional CNN-based approaches.

The second image-based model is a transformer-based object detector. Transformer-based models inherently employ attention mechanisms that emphasize salient regions of the input. The attention is similar to feature selection strategies in terms of identifying important features. As a result, the model achieves high detection accuracy. The improvement is more pronounced in scenarios with limited training data. The attention maps produced by this model are superior to the feature selection modules embedded in CNN architectures.

Not only focusing on the camera, the dissertation also investigates radar signals. This dissertation utilizes data collected from a pulse radar station in Vietnam. Domain knowledge from maritime surveillance experts is incorporated to guide the deep learning model in extracting physically meaningful and task-relevant features. Subsequently, a vessel classification approach based on pulse radar signals is presented. Here, the classification is based on clustering results. Compared to conventional methods, the proposed approach offers three main contributions. First, it introduces a real-world pulse radar dataset acquired from a surveillance station in Vietnam. Unlike previous studies that primarily rely on simulated data, the dataset used in this dissertation is collected under real operating conditions and thus possesses high practical value. Second, expert-driven features derived from maritime surveillance experience are incorporated into the vessel classification process and evaluated through a clustering algorithm. Finally, a radar-based vessel classification algorithm is proposed, in which

vessel categories are determined based on the clustering outcomes of the extracted radar features.

MỞ ĐẦU

Lý do chọn đề tài

Trong công tác giám sát và bảo vệ chủ quyền biển đảo, việc phát hiện, nhận dạng và phân loại các loại tàu biển có ý nghĩa đặc biệt quan trọng, nhất là đối với lực lượng Hải quân và các hệ thống radar cảnh giới ven biển. Hiện nay, các hệ thống giám sát bờ biển chủ yếu dựa trên hai loại cảm biến chính là radar và camera giám sát.

Radar đóng vai trò chủ lực trong việc phát hiện và theo dõi mục tiêu ở cự ly xa, hoạt động ổn định trong mọi điều kiện thời tiết và thời gian ngày đêm. Radar cung cấp các thông tin quan trọng như vị trí, tốc độ và hướng di chuyển của mục tiêu, cho phép phát hiện sớm các nguy cơ va chạm, xâm nhập trái phép hoặc hoạt động bất thường trong vùng biển quản lý. Tuy nhiên, tín hiệu radar chỉ phản ánh đặc tính điện từ của mục tiêu, không cung cấp hình ảnh trực quan, do đó gặp nhiều khó khăn trong việc phân biệt chính xác loại tàu và mục đích hoạt động.

Ngược lại, hệ thống camera giám sát, bao gồm camera quang học và camera hồng ngoại, cung cấp hình ảnh trực quan giúp nhận dạng chi tiết hình dáng, kích thước, cờ hiệu và hành vi của tàu thuyền. Camera ngày càng được triển khai rộng rãi tại các trạm radar và trung tâm quan sát ven biển. Tuy nhiên, camera lại bị hạn chế về cự ly quan sát, chịu ảnh hưởng mạnh của điều kiện thời tiết, ánh sáng và tầm nhìn.

Trước những hạn chế riêng lẻ của từng loại cảm biến, xu hướng kết hợp radar và camera trong giám sát bờ biển đã trở nên tất yếu. Radar thực hiện nhiệm vụ phát hiện và chỉ thị mục tiêu từ xa, trong khi camera đảm nhiệm vai trò ghi hình và phân tích chi tiết. Mặc dù vậy, hiện nay việc nhận dạng và phân loại tàu trong các hệ thống radar – camera vẫn chủ yếu dựa vào kinh nghiệm của nhân viên vận hành, dẫn đến tính chủ quan cao và khó đáp ứng yêu cầu giám sát trong môi trường phức tạp, mật độ tàu thuyền lớn.

Vì vậy, việc nghiên cứu ứng dụng trí tuệ nhân tạo nhằm tự động hóa quá trình nhận dạng và phân loại

tàu biển dựa trên dữ liệu radar và ảnh camera là yêu cầu cấp thiết, góp phần nâng cao hiệu quả giám sát bờ biển, bảo đảm an toàn hàng hải và bảo vệ chủ quyền quốc gia.

Mục đích nghiên cứu

Mục đích của đề tài là nghiên cứu, phát triển phương pháp ứng dụng trí tuệ nhân tạo trong việc nhận dạng và phân loại tàu biển dựa trên tín hiệu phản xạ radar và dữ liệu ảnh từ camera giám sát, nhằm nâng cao độ chính xác, tính khách quan và mức độ tự động hóa trong công tác cảnh giới bờ biển.

Nhiệm vụ nghiên cứu

Để đạt được mục đích đề ra, đề tài tập trung vào việc thực hiện các nhiệm vụ nghiên cứu sau:

- Nghiên cứu và xử lý dữ liệu ảnh từ hệ thống camera giám sát phục vụ nhận dạng trực quan và hỗ trợ phân loại các loại tàu biển.
- Phân tích đặc điểm tín hiệu phản xạ radar của các loại tàu biển hoạt động trong vùng biển Việt Nam.
- Xây dựng và đánh giá các mô hình trí tuệ nhân tạo phục vụ phân loại tàu biển dựa trên tín hiệu radar.
- Xây dựng và đánh giá các mô hình trí tuệ nhân tạo phục vụ phân loại tàu biển dựa trên tín hiệu ảnh.

Phạm vi nghiên cứu

Phạm vi nghiên cứu của đề tài tập trung vào việc khai thác tín hiệu phản xạ radar và dữ liệu ảnh camera phục vụ bài toán phát hiện tàu biển. Cụ thể, bên cạnh việc sử dụng các cơ sở dữ liệu ảnh có sẵn, điểm nhấn của nghiên cứu là đã tiến hành thu thập và khai thác trực tiếp dữ liệu từ một trạm giám sát dọc bờ biển thực tế tại Việt Nam. Việc sử dụng bộ dữ liệu thực tế này nhằm xây dựng, kiểm chứng mô hình và minh chứng toàn diện hiệu quả thực tiễn của các giải pháp đã đề xuất.

Nghiên cứu chủ yếu dừng ở mức phát triển, huấn luyện và đánh giá các mô hình trí tuệ nhân tạo ở dạng nguyên mẫu (prototype), chưa đi sâu vào triển khai hạ tầng phần cứng trên diện rộng.

Đề tài tập trung vào các vấn đề kỹ thuật liên quan đến xử lý tín hiệu radar, xử lý ảnh và hiệu quả thuật toán, không đi sâu vào các khía cạnh pháp lý, chính sách quản lý hay hợp tác quốc tế trong lĩnh vực giám sát biển.

Hướng tiếp cận và phương pháp nghiên cứu

Hướng tiếp cận

Đề tài tiếp cận theo hướng kết hợp đa nguồn dữ liệu từ radar và camera giám sát nhằm tận dụng ưu điểm của từng loại cảm biến. Radar đảm nhiệm vai trò phát hiện và theo dõi mục tiêu ở cự ly xa trong mọi điều kiện thời tiết, trong khi camera cung cấp hình ảnh trực quan hỗ trợ nhận dạng chi tiết. Trên cơ sở đó, đề tài ứng dụng các kỹ thuật trí tuệ nhân tạo, học máy và học sâu để tự động trích xuất đặc trưng và phân loại các loại tàu biển, giảm sự phụ thuộc vào kinh nghiệm chủ quan của nhân viên vận hành.

Phương pháp nghiên cứu Đối với dữ liệu ảnh camera: áp dụng các mô hình học sâu trong bài toán phát hiện và nhận dạng đối tượng, sử dụng các tập dữ liệu chuẩn về tàu biển để huấn luyện và đánh giá mô hình.

Đối với dữ liệu radar: thực hiện tiền xử lý, lọc nhiễu, trích xuất đặc trưng tín hiệu phản xạ, kết hợp các phương pháp phân cụm và phân loại để xây dựng mô hình nhận dạng tàu biển từ tín hiệu radar.

Đánh giá hệ thống dựa trên các tiêu chí: độ chính xác nhận dạng, khả năng phân loại, tốc độ xử lý và độ ổn định trong môi trường biển phức tạp.

Chương 1

TỔNG QUAN

Chương 1 trình bày bối cảnh và yêu cầu của bài toán giám sát an ninh bờ biển trong điều kiện hoạt động hàng hải ngày càng phức tạp. Vai trò của radar và camera trong hệ thống cảnh giới ven bờ được phân tích, đồng thời chỉ ra những hạn chế khi các cảm biến hoạt động độc lập. Các công trình nghiên cứu trong và ngoài nước liên quan đến phát hiện, nhận dạng và phân loại tàu biển được tổng hợp và đánh giá có hệ thống. Xu hướng hợp nhất đa cảm biến và ứng dụng học sâu trong giám sát hàng hải được làm rõ. Trên cơ sở đó, chương xác định khoảng trống nghiên cứu và định hướng giải quyết của luận án.

1.1 Giới thiệu

Biển và vùng ven biển có vai trò rất quan trọng trong nhiều lĩnh vực như: phát triển kinh tế, quốc phòng và an ninh của Việt Nam nói riêng và nhiều quốc gia trên thế giới nói chung. Hoạt động hàng hải trên biển ngày càng gia tăng cả về số lượng lẫn mức độ phức tạp, bao gồm vận tải thương mại, khai thác thủy sản, du lịch biển cũng như các hoạt động quân sự và bán quân sự. Trong bối cảnh đó, yêu cầu giám sát, quản lý và bảo vệ vùng biển ngày càng đặt ra những thách thức lớn đối với các hệ thống quan sát và cảnh giới truyền thống. Việc phát hiện sớm, theo dõi liên tục, nhận dạng chính xác và phân loại kịp thời các mục tiêu tàu biển trở thành một trong những nhiệm vụ then chốt nhằm bảo đảm an toàn hàng hải, duy trì trật tự trên biển và bảo vệ chủ quyền quốc gia.

Trong các hệ thống giám sát bờ biển hiện nay, radar và camera là hai loại cảm biến được sử dụng phổ biến và đóng vai trò bổ trợ cho nhau. Radar có ưu thế nổi bật trong việc phát hiện và theo dõi mục tiêu ở cự ly xa, hoạt động ổn định trong mọi điều kiện khác nhau như thời tiết và thời gian giám sát. Thông tin do radar cung cấp như vị trí, vận tốc và hướng di chuyển của mục tiêu cho phép xây dựng

bức tranh tổng thể về tình hình hoạt động trên biển trong thời gian thực. Tuy nhiên, tín hiệu radar chủ yếu phản ánh đặc tính phản xạ điện từ của mục tiêu, thiếu thông tin trực quan, do đó gặp nhiều hạn chế trong việc nhận dạng chính xác loại tàu và đánh giá mục đích hoạt động.

Ngược lại, hệ thống camera giám sát, bao gồm camera quang học và camera hồng ngoại, cung cấp hình ảnh trực quan giúp quan sát chi tiết hình dáng, kích thước, cấu trúc và hành vi của tàu thuyền. Nhờ đó, camera hỗ trợ hiệu quả cho công tác nhận dạng và phân loại mục tiêu, đặc biệt trong các tình huống cần xác định rõ loại tàu hoặc phát hiện hành vi bất thường. Tuy nhiên, khả năng hoạt động của camera bị giới hạn bởi cự ly quan sát, điều kiện ánh sáng và thời tiết, khiến hiệu quả giám sát suy giảm trong môi trường biển phức tạp như sương mù, mưa lớn hoặc ban đêm.

Trước những giới hạn của từng loại cảm biến riêng lẻ, việc tích hợp radar và camera trong hệ thống giám sát bờ biển đã trở thành xu hướng tất yếu. Radar đảm nhiệm vai trò phát hiện và chỉ thị mục tiêu từ xa, trong khi camera được điều khiển hướng về vị trí mục tiêu để ghi hình và phân tích chi tiết. Cách tiếp cận này cho phép tận dụng tối đa ưu điểm của cả hai loại cảm biến, góp phần nâng cao hiệu quả giám sát tổng thể. Tuy nhiên, trên thực tế, quá trình nhận dạng và phân loại tàu trong các hệ thống radar–camera hiện nay vẫn phụ thuộc nhiều vào kinh nghiệm và khả năng đánh giá chủ quan của nhân viên vận hành. Điều này không chỉ làm giảm tính nhất quán của kết quả phân loại mà còn gây khó khăn trong việc xử lý khối lượng dữ liệu lớn và yêu cầu phản ứng nhanh trong các tình huống phức tạp.

Sự phát triển mạnh mẽ của trí tuệ nhân tạo trong những năm gần đây đã mở ra nhiều hướng tiếp cận mới cho các bài toán nhận dạng và phân loại mục tiêu. Các phương pháp học máy và học sâu cho thấy khả năng vượt trội trong việc xử lý dữ liệu lớn, tự động trích xuất đặc trưng và học các mẫu phức tạp từ dữ liệu đa chiều. Trong lĩnh vực xử lý ảnh, các mô hình học sâu đã đạt được nhiều thành tựu nổi bật trong bài toán phát hiện và nhận dạng đối tượng, bao gồm cả các mục tiêu hàng hải. Trong khi đó, các phương pháp học máy cũng được ứng dụng ngày càng rộng rãi trong xử lý tín hiệu radar, cho phép khai thác hiệu quả các đặc trưng phản xạ phục vụ cho bài toán phân loại mục tiêu.

Mặc dù đã có nhiều nghiên cứu tập trung vào nhận dạng tàu biển từ ảnh camera hoặc phân loại mục tiêu từ tín hiệu radar, phần lớn các công trình vẫn xem xét hai loại dữ liệu này một cách tách biệt. Việc kết hợp có hệ thống giữa tín hiệu radar và dữ liệu ảnh camera, đặc biệt trong bối cảnh giám sát bờ biển, vẫn còn nhiều vấn đề chưa được nghiên cứu đầy đủ. Bên cạnh đó, các trạm radar thực tế thường phải đối mặt với dữ liệu không đồng nhất, nhiễu cao và điều kiện quan sát thay đổi liên tục, đòi hỏi các phương pháp xử lý và mô hình phân loại có độ tin cậy và khả năng thích nghi cao.

Xuất phát từ những yêu cầu thực tiễn và khoảng trống nghiên cứu nêu trên, luận án này tập trung nghiên cứu bài toán nhận dạng và phân loại tàu biển trong hệ thống cảnh giới bờ biển dựa trên sự kết hợp giữa tín hiệu phản xạ radar và dữ liệu ảnh camera, với sự hỗ trợ của các phương pháp trí tuệ nhân tạo. Luận án hướng tới xây dựng các mô hình có khả năng tự động trích xuất đặc trưng từ dữ liệu radar và ảnh, học mối quan hệ giữa các đặc trưng này và đưa ra quyết định phân loại một cách khách quan, nhất quán và hiệu quả.

Bên cạnh việc nâng cao độ chính xác trong nhận dạng và phân loại mục tiêu, nghiên cứu còn hướng tới việc giảm sự phụ thuộc vào yếu tố con người trong quá trình vận hành hệ thống giám sát. Thông qua việc tự động hóa các khâu phân tích và đánh giá, hệ thống đề xuất có thể hỗ trợ nhân viên radar trong việc ra quyết định, đặc biệt trong các tình huống yêu cầu xử lý nhanh và chính xác. Đồng thời, kết quả nghiên cứu cũng góp phần tạo nền tảng khoa học cho việc phát triển các hệ thống cảnh giới bờ biển thông minh, có khả năng mở rộng và thích nghi với các điều kiện hoạt động khác nhau.

Với cách tiếp cận kết hợp giữa tri thức chuyên gia và các mô hình trí tuệ nhân tạo hiện đại, luận án kỳ vọng đóng góp cả về mặt lý thuyết và thực tiễn trong lĩnh vực giám sát bờ biển. Những kết quả đạt được không chỉ góp phần hoàn thiện các phương pháp nhận dạng và phân loại tàu biển mà còn có tiềm năng ứng dụng rộng rãi trong các hệ thống quan sát hàng hải, góp phần bảo đảm an toàn, an ninh và phát triển bền vững các hoạt động trên biển.

1.2 Các công bố quốc tế về giám sát an ninh bờ biển

Giám sát hoạt động tàu thuyền ven bờ là một nhiệm vụ quan trọng trong quản lý tài nguyên biển và thực thi pháp luật hàng hải. Các nghiên cứu ban đầu tập trung vào việc khai thác radar ven bờ (coastal radar) như một công cụ giám sát liên tục, độc lập với điều kiện ánh sáng và thời tiết. Công trình của Cope et al. (2022) [3] đã chứng minh khả năng sử dụng radar X-band thương mại để theo dõi hoạt động tàu thuyền ở quy mô nhỏ trong các khu bảo tồn biển (MPAs), với độ phân giải không gian–thời gian cao và khả năng vận hành liên tục trong thời gian dài. Kết quả cho thấy radar không chỉ phát hiện và bám vết mục tiêu hiệu quả mà còn cho phép suy luận hành vi (ví dụ: hoạt động đánh bắt) thông qua phân tích quỹ đạo chuyển động. Tuy nhiên, nghiên cứu này cũng chỉ ra hạn chế cố hữu của radar đơn lẻ, đặc biệt là khả năng nhận dạng và phân loại mục tiêu còn hạn chế, từ đó đặt ra nhu cầu kết hợp thêm các cảm biến bổ sung nhằm nâng cao mức độ hiểu biết ngữ cảnh.

Song song với các ứng dụng radar giám sát tổng quát, một hướng nghiên cứu khác tập trung vào radar

xung (pulse radar) và radar Pulse-Doppler [4] cho bài toán nhận dạng mục tiêu tự động (ATR) trong cảnh giới ven bờ. Việc sử dụng radar xung cho phép giám sát ở quy mô rộng và cự li xa. Do đó phù hợp cho các vấn đề trong quân sự. Công trình kinh điển về ATR cho radar Pulse-Doppler đã đề xuất việc khai thác các đặc trưng tín hiệu như Doppler, RCS và các mô hình chuyển động để phân loại tàu cá trong môi trường thực. Việc kết hợp các đặc trưng vật lý với thuật toán phân loại cho thấy hiệu quả nhất định, đặc biệt trong điều kiện thiếu dữ liệu hình ảnh. Tuy nhiên, các phương pháp này thường yêu cầu thiết kế đặc trưng thủ công, nhạy cảm với nhiễu và khó mở rộng khi số lớp mục tiêu tăng lên. Điều này thúc đẩy xu hướng tích hợp radar với các cảm biến giàu thông tin ngữ nghĩa hơn, đặc biệt là camera quang học.

Trong những năm gần đây, camera RGB kết hợp với kỹ thuật học sâu đã trở thành công cụ chủ đạo cho bài toán phát hiện và phân loại tàu biển. Bài tổng quan của Yang et al. (2024) [5] hệ thống hóa các phương pháp nhận dạng mục tiêu hàng hải dựa trên ảnh RGB, từ các mô hình CNN truyền thống đến các kiến trúc hiện đại như YOLO, Faster R-CNN và Transformer. Quan trọng hơn, camera đơn lẻ khó cung cấp thông tin vị trí và vận tốc chính xác, làm hạn chế khả năng theo dõi và phân tích hành vi mục tiêu trong thời gian dài.

Bên cạnh các nghiên cứu mang tính thuật toán, một số công trình đã triển khai hệ thống camera + AI trong các kịch bản giám sát ven bờ thực tế. Nghiên cứu [6] (2024) trình bày một hệ thống giám sát tàu cá ven bờ dựa trên kỹ thuật hoc sâu và hạ tầng đám mây, cho phép tự động nhận dạng hàng trăm loài sinh vật biển và trích xuất các chỉ số phục vụ quản lý nghề cá. Mặc dù trọng tâm không phải giám sát tàu thuyền, công trình này minh chứng tính khả thi và hiệu quả của camera kết hợp AI trong môi trường ven bờ quy mô lớn. Tuy nhiên, hệ thống vẫn phụ thuộc mạnh vào điều kiện quan sát quang học, cho thấy sự cần thiết của việc tích hợp thêm radar để đảm bảo tính liên tục và độ tin cậy.

Nhằm khắc phục các hạn chế của từng cảm biến đơn lẻ, hợp nhất dữ liệu đa cảm biến đã trở thành hướng nghiên cứu trung tâm trong giám sát hàng hải. Công trình kinh điển của Molina et al. [7] đã đặt nền móng cho các hệ thống fusion radar–camera–AIS, nhấn mạnh các vấn đề thực tế như tính bền vững, khả năng mở rộng và kết hợp dữ liệu trong môi trường mật độ cao. Trên cơ sở đó, báo cáo của của Zhou et al. (2022) [8] đã hệ thống hóa các kỹ thuật fusion radar mmWave và camera, bao gồm hiệu chuẩn cảm biến, căn chỉnh không–thời gian và các mức fusion (data-level, feature-level, decision-level). Bài báo này cung cấp cơ sở quan trọng cho việc thiết kế quy tình xử lý hợp nhất cảm biến trong các hệ thống giám sát ven bờ hiện đại.

Gần đây, sự phát triển của deep learning đã thúc đẩy mạnh mẽ các phương pháp deep fusion radar và

camera. Bài nghiên cứu [9] đã cung cấp tổng quan toàn diện về các kiến trúc fusion sâu cho bài toán phát hiện và theo dõi mục tiêu, bao gồm biểu diễn đa modal, căn chỉnh dữ liệu và chiến lược fusion trong mạng học sâu. Các phương pháp này đặc biệt phù hợp với môi trường ven biển phức tạp, nơi radar đảm bảo khả năng phát hiện trong điều kiện xấu, còn camera cung cấp thông tin ngữ nghĩa chi tiết. Tuy nhiên, survey cũng chỉ ra những thách thức mở, đặc biệt là thiếu dataset đa cảm biến chuẩn hóa và khó khăn trong việc đồng bộ hóa radar-camera ngoài thực địa.

Các nghiên cứu học sâu thường đòi hỏi các tập dữ liệu hàng hải đa cảm biến. Picoastal [3] nỗ lực thu hẹp khoảng cách giữa nghiên cứu và triển khai thực tế bằng cách thiết kế một hệ thống đa cảm biến chi phí thấp. Tuy nhiên, các nghiên cứu cũng chỉ ra rằng dữ liệu kết hợp radar xung/Pulse-Doppler và camera ven bờ vẫn còn rất hạn chế, đặc biệt là thiếu ground truth đồng bộ. Khoảng trống này mở ra hướng nghiên cứu quan trọng cho các hệ thống giám sát ven bờ thế hệ mới, nơi việc kết hợp radar xung và camera cùng các thuật toán AI đóng vai trò then chốt.

1.3 Các công bố trong nước về giám sát an ninh bờ biển

Ở Việt Nam đa số các công trình nghiên cứu tập trung vào HF-radar cho nghiên cứu hải dương (dòng bề mặt, xói lở, mô tả thủy động lực). Về lĩnh vực radar Pulse / Pulse-Doppler phục vụ ATR (nhận dạng mục tiêu tàu) — công bố chuyên sâu từ nhóm tác giả Việt Nam còn hạn chế. Trong lĩnh vực ứng dụng camera vào giám sát tàu biển, nhiều công trình luận án trong nước đã đề cập tới nhưng còn ít bộ dữ liệu công khai lớn và rất ít công trình công bố quốc tế kết hợp đồng bộ radar xung + camera phục vụ ATR cho giám sát ven bờ.

Một số nghiên cứu tiêu biểu ở VN về vấn đề giám sát bờ biển có thể liệt kê như sau.

Nghiên cứu của tác giả Trần Thanh Huyền [10]là một trong những công trình đầu tiên khai thác dữ liệu dòng chảy bề mặt độ phân giải cao từ hệ thống HF radar tại khu vực ven biển Nam Trung Bộ Việt Nam trong giai đoạn chuyển mùa gió mùa hè châu Á. Kết quả cho thấy hoàn lưu bề mặt tại khu vực này có tính biến thiên mạnh theo không gian và thời gian, chịu ảnh hưởng đồng thời của gió mùa và các yếu tố động lực nội tại của đại dương. Việc so sánh giữa dữ liệu đo HF radar và mô phỏng từ mô hình SYMPHONIE chỉ ra sự sai khác đáng kể trong một số thời điểm, đặc biệt khi trường gió biến đổi. Hai phương pháp tối ưu hóa trường gió (EnPS và EkW) đã giúp giảm sai số vận tốc dòng chảy bề mặt khoảng 36–40% so với dữ liệu đo, khẳng định vai trò quan trọng của dữ liệu radar ven bờ trong hiệu chỉnh mô hình số và cải thiện chất lượng dự báo. Nghiên cứu cũng chỉ ra rằng hoàn lưu bề mặt

không chỉ phụ thuộc vào gió mà còn chịu tác động của biến thiên nội tại đại dương, gợi mở hướng tiếp cận tích hợp dữ liệu lớn và hợp nhất dữ liệu (data fusion) trong các bài toán giám sát biển.

Ở góc độ chiến lược và chính sách, tài liệu tác giả Bích Tran [11] nhấn mạnh vai trò then chốt của nhận thức tình huống hàng hải đối với bảo vệ chủ quyền, an toàn hàng hải và lợi ích kinh tế biển của Việt Nam. Báo cáo chỉ ra rằng Việt Nam đang từng bước hiện đại hóa năng lực giám sát dưới nước, trên mặt biển và ven bờ thông qua đầu tư vào các nền tảng trên biển, trên không và không gian, kết hợp với hợp tác quốc tế và chia sẻ thông tin. Tuy nhiên, để đáp ứng yêu cầu giám sát một vùng biển rộng lớn và đường bờ biển dài, Việt Nam cần đẩy mạnh ứng dụng các công nghệ tiên tiến, đặc biệt là công nghệ không gian và các giải pháp giám sát ven bờ có tính tự động cao. Nhận định này cho thấy nhu cầu cấp thiết trong việc nghiên cứu, phát triển các hệ thống giám sát thông minh dựa trên radar, cảm biến và trí tuệ nhân tạo.

Trong lĩnh vực đo đạc thực nghiệm, công trình của N. K. Cường [12] đã cung cấp bộ dữ liệu đo sóng và dòng chảy bề mặt có độ phân giải cao tại vùng biển ngoài khơi tỉnh Bình Thuận – khu vực có tiềm năng năng lượng gió lớn nhất Việt Nam. Việc lắp đặt hai hệ thống HF radar trên bờ cho phép thu thập dữ liệu với độ phân giải không gian 1,5 km và chu kỳ thời gian 30 phút trên vùng phủ lên tới 30 km × 50 km. Kết quả đo đạc phản ánh rõ sự khác biệt giữa hai mùa gió mùa Đông Bắc và Tây Nam, với cường độ sóng, gió và dòng chảy biến đổi mạnh theo mùa. Nghiên cứu khẳng định giá trị của dữ liệu HF radar trong quan trắc biển ven bờ, phục vụ không chỉ cho nghiên cứu khoa học mà còn cho khai thác tài nguyên và đánh giá tiềm năng năng lượng tái tạo.

Bên cạnh đó, nghiên cứu của N.Mai [13] tập trung vào bài toán cải thiện chất lượng dữ liệu dòng chảy bề mặt thu từ HF radar thông qua các phương pháp nội suy và đồng hóa dữ liệu. Do dữ liệu radar thường bị gián đoạn theo không gian và thời gian bởi nhiễu tín hiệu và điều kiện biển, tác giả đã áp dụng phương pháp nội suy dựa trên hàm trực giao thực nghiệm (EOF) kết hợp với kỹ thuật nội suy biến phân hai chiều (2dVar). Kết quả thử nghiệm ban đầu trên dữ liệu HF radar tại Việt Nam cho thấy các phương pháp này có khả năng cải thiện đáng kể tính liên tục và độ tin cậy của trường dòng chảy bề mặt, mở ra hướng tiếp cận hiệu quả trong xử lý và khai thác dữ liệu radar ven bờ.

Tổng hợp các nghiên cứu trên cho thấy HF radar là công cụ quan trọng trong giám sát biển ven bờ, từ đo đạc thực nghiệm, cải thiện mô hình số đến nâng cao nhận thức tình huống hàng hải. Tuy nhiên, phần lớn các nghiên cứu hiện nay vẫn tập trung vào mô tả động lực học biển hoặc cải thiện chất lượng dữ liệu, trong khi việc khai thác dữ liệu radar cho bài toán nhận dạng, phân loại mục tiêu và hỗ trợ cảnh giới tự động vẫn còn nhiều dư địa nghiên cứu, đặc biệt khi kết hợp với các phương pháp trí tuệ

nhân tạo.

1.4 Kết luận của chương

Qua tổng hợp và phân tích các công bố khoa học trong và ngoài nước, có thể khẳng định rằng giám sát an ninh bờ biển là một bài toán phức tạp, mang tính liên ngành cao, chịu tác động đồng thời của các yếu tố kỹ thuật, môi trường và yêu cầu thực tiễn. Sự gia tăng nhanh chóng của các hoạt động hàng hải, từ vận tải thương mại, khai thác thủy sản đến các hoạt động quân sự và bán quân sự, đã đặt ra yêu cầu ngày càng cao đối với các hệ thống giám sát ven bờ, không chỉ dừng ở phát hiện và theo dõi mục tiêu mà còn phải hướng tới nhận dạng, phân loại và đánh giá hành vi một cách kịp thời và chính xác.

Các nghiên cứu quốc tế cho thấy radar ven bờ đóng vai trò nền tảng trong các hệ thống giám sát hàng hải hiện đại. Với khả năng hoạt động liên tục, độc lập với điều kiện ánh sáng và thời tiết, radar – đặc biệt là radar X-band, radar xung và Pulse-Doppler – cho phép phát hiện và theo dõi mục tiêu ở cự ly xa, cung cấp các thông tin định lượng quan trọng như vị trí, vận tốc và quỹ đạo chuyển động. Một số công trình đã khai thác các đặc trưng vật lý của tín hiệu radar, bao gồm Doppler, tiết diện phản xạ radar (RCS) và các mô hình chuyển động, để phục vụ bài toán nhận dạng mục tiêu tự động. Tuy nhiên, các phương pháp này phần lớn vẫn dựa vào thiết kế đặc trưng thủ công, nhạy cảm với nhiễu môi trường và khó mở rộng khi số lượng lớp mục tiêu tăng lên hoặc khi điều kiện hoạt động thay đổi.

Ở chiều ngược lại, các nghiên cứu dựa trên camera quang học kết hợp với các mô hình học sâu đã đạt được những kết quả ấn tượng trong bài toán phát hiện và phân loại tàu biển. Dữ liệu hình ảnh cung cấp thông tin trực quan giàu ngữ nghĩa, cho phép nhận dạng chi tiết hình dáng, kích thước và cấu trúc mục tiêu – những yếu tố mà radar khó hoặc không thể cung cấp. Sự phát triển nhanh chóng của các kiến trúc mạng học sâu, từ CNN truyền thống đến các mô hình một giai đoạn, hai giai đoạn và Transformer, đã góp phần nâng cao đáng kể độ chính xác và khả năng khái quát của các hệ thống dựa trên camera. Tuy nhiên, hiệu quả của camera lại phụ thuộc mạnh vào điều kiện quan sát, bị giới hạn bởi cự ly, ánh sáng và thời tiết, khiến việc giám sát liên tục và tin cậy trong môi trường biển thực tế vẫn còn nhiều thách thức.

Những hạn chế mang tính bổ sung của radar và camera đã thúc đẩy mạnh mẽ xu hướng hợp nhất dữ liệu đa cảm biến trong giám sát bờ biển. Các nghiên cứu về kết hợp cả hai loại cảm biến nhằm phân loại tàu biển cho thấy việc kết hợp nhiều nguồn thông tin có thể cải thiện đáng kể mức độ nhận thức tình huống hàng hải, giảm thiểu các điểm mù và nâng cao độ tin cậy của hệ thống. Đặc biệt, sự phát

triển của học sâu đa modal và các kiến trúc deep fusion đã mở ra khả năng tự động học các mối quan hệ phức tạp giữa dữ liệu radar và dữ liệu ảnh, thay thế dần các chiến lược fusion thủ công truyền thống. Mặc dù vậy, các công trình tổng quan gần đây cũng chỉ ra rằng việc triển khai các phương pháp fusion trong môi trường ven bờ thực tế vẫn gặp nhiều rào cản, bao gồm vấn đề đồng bộ không–thời gian giữa các cảm biến, nhiễu tín hiệu, sự không đồng nhất của dữ liệu và sự thiếu hụt các bộ dữ liệu đa cảm biến được gán nhãn đầy đủ.

Trong bối cảnh nghiên cứu trong nước, các công trình khoa học tại Việt Nam đã đạt được những kết quả đáng kể trong việc ứng dụng HF radar cho quan trắc biển ven bờ, đặc biệt là đo đạc dòng chảy bề mặt, sóng biển và hỗ trợ mô hình số hải dương. Những nghiên cứu này không chỉ góp phần nâng cao hiểu biết về động lực học biển mà còn khẳng định giá trị của radar ven bờ trong việc cung cấp dữ liệu quan sát có độ phân giải cao và liên tục. Ở góc độ chiến lược, các báo cáo và nghiên cứu chính sách cũng nhấn mạnh tầm quan trọng của việc nâng cao nhận thức tình huống hàng hải đối với bảo đảm chủ quyền, an toàn và phát triển kinh tế biển của Việt Nam.

Tuy nhiên, khi xem xét một cách hệ thống, có thể nhận thấy rằng các nghiên cứu trong nước về nhận dạng và phân loại các loại tàu biển, đặc biệt là sử dụng radar xung hoặc Pulse-Doppler kết hợp với camera và trí tuệ nhân tạo, vẫn còn rất hạn chế. Phần lớn các công trình mới tập trung vào mô tả hiện tượng hải dương hoặc cải thiện chất lượng dữ liệu radar, trong khi các bài toán cảnh giới tự động, nhận dạng mục tiêu và hỗ trợ ra quyết định vẫn chưa được nghiên cứu đầy đủ. Bên cạnh đó, việc thiếu các bộ dữ liệu đa cảm biến được thu thập trong điều kiện thực tế và được gán nhãn đồng bộ cũng là một rào cản lớn đối với việc phát triển và đánh giá các mô hình học máy và học sâu trong bối cảnh Việt Nam.

Tổng hợp các phân tích từ chương này cho thấy, mặc dù các công nghệ giám sát đơn lẻ như radar hay camera đã đạt được những kết quả đáng ghi nhận, chúng vẫn bộc lộ những hạn chế cố hữu khi hoạt động độc lập. Radar có lợi thế giám sát tầm xa nhưng thiếu thông tin chi tiết để định danh mục tiêu, trong khi camera cung cấp hình ảnh trực quan nhưng lại phụ thuộc lớn vào điều kiện thời tiết và ánh sáng. Thực tiễn an ninh hàng hải hiện đại đòi hỏi một hệ thống có khả năng khắc phục các nhược điểm này để hoạt động bền bỉ và chính xác trong mọi tình huống.

Tuy nhiên, các nghiên cứu trong nước hiện nay phần lớn vẫn tập trung vào xử lý riêng lẻ từng loại cảm biến mà thiếu đi các cơ chế phối hợp hoạt động thông minh giữa chúng. Chính vì vậy, việc nghiên cứu ứng dụng các thuật toán học sâu tiên tiến để tối ưu hóa năng lực xử lý nhận dạng hình ảnh và tín hiệu, đồng thời xây dựng cơ chế điều khiển phối hợp giữa radar và camera là một yêu cầu khoa học cấp

thiết. Luận án này sẽ tập trung giải quyết vấn đề trên nhằm đề xuất một hệ thống cảnh giới bờ biển có khả năng tự động hóa cao, trong đó radar đóng vai trò dẫn đường và camera thực hiện nhận dạng chi tiết, phù hợp với điều kiện thực tế tại Việt Nam.

Chương 2

CƠ SỞ LÝ THUYẾT

Chương 2 trình bày nền tảng lý thuyết phục vụ cho các phương pháp đề xuất trong luận án. Trước hết, cơ sở lý thuyết về phát hiện đối tượng trong ảnh, bao gồm các mô hình học sâu hiện đại và các chỉ số đánh giá hiệu năng, được phân tích chi tiết. Tiếp theo, các phương pháp phân cụm và phân loại dữ liệu, đặc biệt trong bối cảnh xử lý tín hiệu radar, được hệ thống hóa. Các khái niệm về trích xuất đặc trưng, học biểu diễn và tối ưu hóa mô hình được trình bày nhằm làm rõ cơ sở khoa học cho các cải tiến sau này. Chương này đóng vai trò nền tảng lý luận cho các nghiên cứu thực nghiệm ở các chương tiếp theo.

2.1 Cơ sở lý thuyết về phát hiện đối tượng bằng cảm biến ảnh

2.1.1 Hệ thống camera giám sát bờ biển

Một cách truyền thống, bờ biển được giám sát bằng các hệ thống radar. Các hệ thống radar giám sát bờ biển có rất nhiều ưu điểm như có thể phát hiện được mục tiêu ở cự ly xa, ít bị ảnh hưởng do điều kiện thời tiết mây mưa, ngày đêm. Tuy nhiên cũng có một số nhược điểm như không thể biết chính xác loại tàu, khi hai mục tiêu tàu biển ở sát nhau không nằm trong khả năng phân biệt theo cự ly và phương vị của radar thì hệ thống sẽ nhận biết đó là một mục tiêu. Chính vì vậy hệ thống camera được phát triển thêm trên các trạm radar cảnh giới bờ biển hỗ trợ cho radar. Một số loại camera thường được sử dụng trên các trạm radar cảnh giới bờ biển như Camera Q6215-LE, và Wisenet. Một số yêu cầu kỹ thuật của camera giám sát bờ biển được liệt kê như sau:

- Camera hoạt động tốt trong điều kiện khắc nghiệt
- Tích hợp chiếu sáng hồng ngoại 400M (1300ft),

Hình 2.1: Cấu tạo của thiết bị quang điện tử. (1) Khối Camera nhiệt; (2) Khối Camera màu; (3) Khối chân đế và khối quay; (4) Khối nguồn

- Khả năng chống nước IP66
- Cảm biến ½ với HDTV 1080p, zoom quang 30x, 60 khung Hình/giây.
- Hỗ trợ chuẩn H264 với Zipstream và Motion JPEG, WDR.

Camera quan sát cảng biển chuyên dụng của hãng Q6215-LE là camera quan sát cảng biển thông minh với tầm nhìn bán kính lên tới 15 km, được thiết kế đặc biệt với khả năng quét, nghiêng và thu phóng chính xác cao, cùng với IR tầm xa để bao quát các khoảng cách rộng và xa trong bán kính 15 km. Máy ảnh này có thể nhận dạng và xác định mục tiêu trong các khu vực mở rộng ngay cả trong điều kiện ánh sáng yếu hoặc bóng tối hoàn toàn. Từ bến cảng biển, cảng hàng không đến đường cao tốc. Đây là camera lý tưởng cho việc giám sát 24/7 trong mọi thời tiết khác nhau.

Camera Q6215-LE chuyên dụng quan sát biển này đáp ứng tiêu chuẩn MIL-STD-810G và NEMA TS-2, (chống ăn mòn, cường độ gió...) bảo đảm hoạt động đáng tin cậy. Có thể hoạt động ổn định vào điều kiện thời tiết khắc nhiệt và có thể chịu được sức gió lên đến 245 km/h (152 li/giờ). Các camera này có vỏ được thiết kế đạt tiêu chuẩn chống va đập IK10 và chống nước IP66/IP68.

Trên các hệ thống camera chuyên dụng cho ứng dụng giám sát bờ biển, người ta thường tích hợp các

bộ đo khoảng cách dựa trên cảm biến laser để xác định khoảng cách đến mục tiêu, cùng với GPS và la bàn từ tính kỹ thuật số để xác định hướng. Ngoài hệ thống camera RGB, các hệ thống này còn trang bị camera nhiệt độ nét cao để quan sát trong đêm.

Tất cả các biến cảm biến của hệ thống được gắn trên khối quay cơ học có thể điều khiển được để điều khiển hướng quan sát. Máy ảnh có khả năng thu phóng liên tục để xác định và nhận dạng mục tiêu. Cũng như có thể hiển thị camera nhiệt độ và camera màu đồng thời. Thiết bị có khả năng hoạt động trong nhiều loại điều kiện thời tiết để đảm bảo việc quan sát không bị gián đoạn và yêu cầu nhiệm vụ ở tần suất cao.

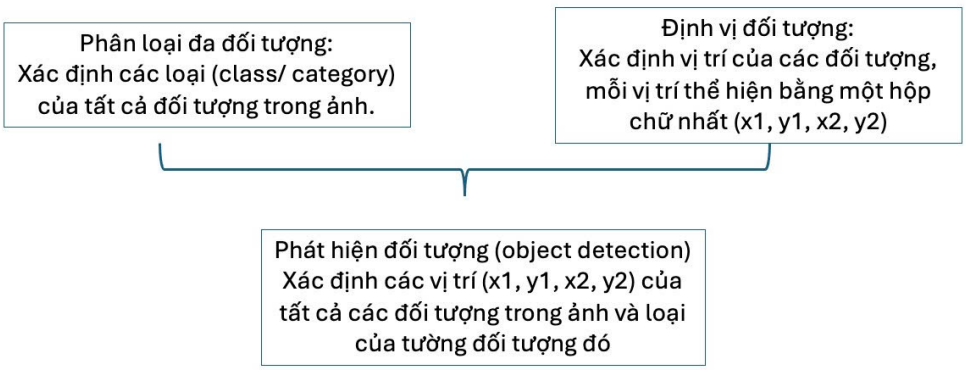
Trong thực tế, việc quan sát và phát hiện mục tiêu của camera có thể đạt được từ 10 đến 20 hải lý. Tuy nhiên nhược điểm của việc quan sát, phát hiện bằng camera cũng có một số hạn chế như đòi hỏi người dùng phải thao tác thủ công trong việc zoom, bắt đối tượng và hệ thống bị ảnh hưởng nhiều của điều kiện thời tiết như mây mưa, sương mù, ban đêm.

2.1.2 Phát hiện đối tượng trong ảnh

Phát hiện đối tượng hay object detection là một kỹ thuật thị giác máy tính hoạt động để xác định và định vị một hoặc nhiều đối tượng thuộc các lớp (class) nhất định trong một Hình ảnh hoặc video. Cụ thể, phát hiện đối tượng vẽ các đường bao (hộp giới hạn) xung quanh các đối tượng được phát hiện này và gán cho chúng một nhãn. Khái niệm phát hiện đối tượng thường hay bị nhầm lẫn với hai khái niệm phân loại hình ảnh (Image Classification) và định vị đối tượng (Object Localization).

- phân loại hình ảnh: dự đoán nhãn của một đối tượng trong Hình ảnh với input là một Hình ảnh chứa một đối tượng và output trả về là nhãn lớp của đối tượng đó.
- Định vị đối tượng: Xác định vị trí hiện diện của các đối tượng trong ảnh và cho biết vị trí của chúng bằng đường bao. Input của nó là một Hình ảnh chứa một hoặc nhiều đối tượng và output trả về là một hoặc nhiều đường bao được xác định bởi tọa độ tâm, chiều rộng và chiều cao.
- Phát hiện đối tượng: Xác định vị trí hiện diện của các đối tượng bằng các đường bao và xác định nhãn của các đối tượng xuất hiện trong Hình ảnh. Đầu vào của nó là một Hình ảnh chứa một hoặc nhiều đối tượng và output trả về một hoặc nhiều đường bao và nhãn cho mỗi đường bao.

Có thể thấy phát hiện đối tượng là một khía cạnh nổi bật của thị giá máy tính. Nó chính là sự kết hợp giữa cả hai kỹ thuật là phân loại hình ảnh và định vị đối tượng.



Hình 2.2: Mối quan hệ giữa các bài toán nhận dạng đối tượng trong thị giác máy tính

Các phương pháp phát hiện đối tượng có thể được phân loại thành hai nhóm tiếp cận chính: (i) các phương pháp truyền thống dựa trên học máy, trong đó sử dụng kỹ thuật trích xuất đặc trưng thủ công kết hợp với các bộ phân loại cổ điển; và (ii) các phương pháp hiện đại dựa trên học sâu, cho phép mô hình tự động trích xuất và học các đặc trưng phức tạp thông qua kiến trúc mạng nơ-ron sâu.

2.1.3 Phát hiện đối tượng bằng đặc trưng thủ công

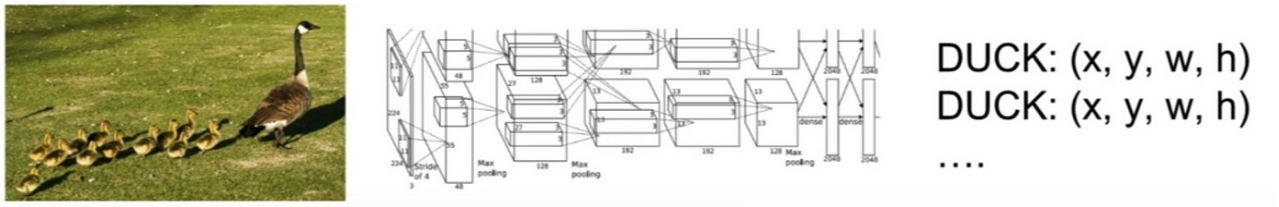
Trước khi mô hình phát hiện đối tượng bằng học sâu được phát triển, việc phát hiện đối tượng được xây dựng dựa trên các đặc trưng thủ công và kiến trúc mô hình nông. Các bước thực hiện của các mô hình phát hiện đối tượng truyền thống chủ yếu có thể được chia thành ba giai đoạn: lựa chọn khu vực chứa thông tin đối tượng, trích xuất đặc trưng và phân loại.

Vì các đối tượng khác nhau có thể xuất hiện ở bất kỳ vị trí nào của Hình ảnh và có tỷ lệ hoặc kích thước khác nhau, nên có thể sử dụng việc quét toàn bộ Hình ảnh bằng cửa sổ trượt đa tỷ lệ để lựa chọn khu vực chứa thông tin đối tượng. Tuy nhiên, với số lượng lớn các cửa sổ được đề xuất, việc này tốn kém về mặt tính toán và tạo ra quá nhiều cửa sổ dư thừa. Nhưng nếu chỉ áp dụng một số lượng mẫu cửa sổ trượt cố định, thì có thể tạo ra những vùng không đạt yêu cầu. Tiếp theo, các mô hình phát hiện đối tượng truyền thống sẽ tiến hành trích xuất đặc trưng của các đối tượng dựa trên các bộ trích xuất như SIFT hay HOG và sử dụng các thuật toán học máy như SVM hay DPM để phân loại đối tượng. Tuy nhiên, các đặc trưng được cung cấp bởi các bộ trích xuất này là những đặc trưng cấp thấp, chỉ là những đặc trưng trên bề mặt nổi của Hình ảnh. Với tốc độ phát triển như hiện nay, dữ liệu của chúng ta ngày càng nhiều và các bài toán bắt đầu khó dần lên đòi hỏi phải trích xuất những đặc trưng cấp cao hơn và được trích xuất sâu hơn dẫn đến những mô hình phát hiện đối tượng truyền thống trở nên kém hiệu quả.

2.1.4 Phát hiện đối tượng dựa trên mô hình học sâu

Thị giác máy tính là một lĩnh vực liên ngành đã đạt được những bước tiến vượt bậc trong những năm gần đây, đặc biệt là nhờ sự phát triển của các Mạng thần kinh tích chập. Trong đó, bài toán phát hiện đối tượng đóng vai trò nền tảng, hỗ trợ đắc lực cho nhiều ứng dụng phức tạp như hệ thống xe tự lái, ước lượng tư thế, giám sát an ninh và phân tích giao thông.

Khác với việc chỉ xử lý một đối tượng duy nhất, bài toán phát hiện đa đối tượng trong môi trường thực tế yêu cầu hệ thống phải dự đoán đồng thời vị trí và nhận dạng các đối tượng xuất hiện trong ảnh. Như được minh họa trong Hình 2.3, số lượng đối tượng trong mỗi bức ảnh là không cố định. Chính đặc điểm đầu ra có chiều dài biến thiên này làm cho các kiến trúc CNN tiêu chuẩn kết nối với các lớp fully-connected truyền thống không thể áp dụng trực tiếp để giải quyết vấn đề.



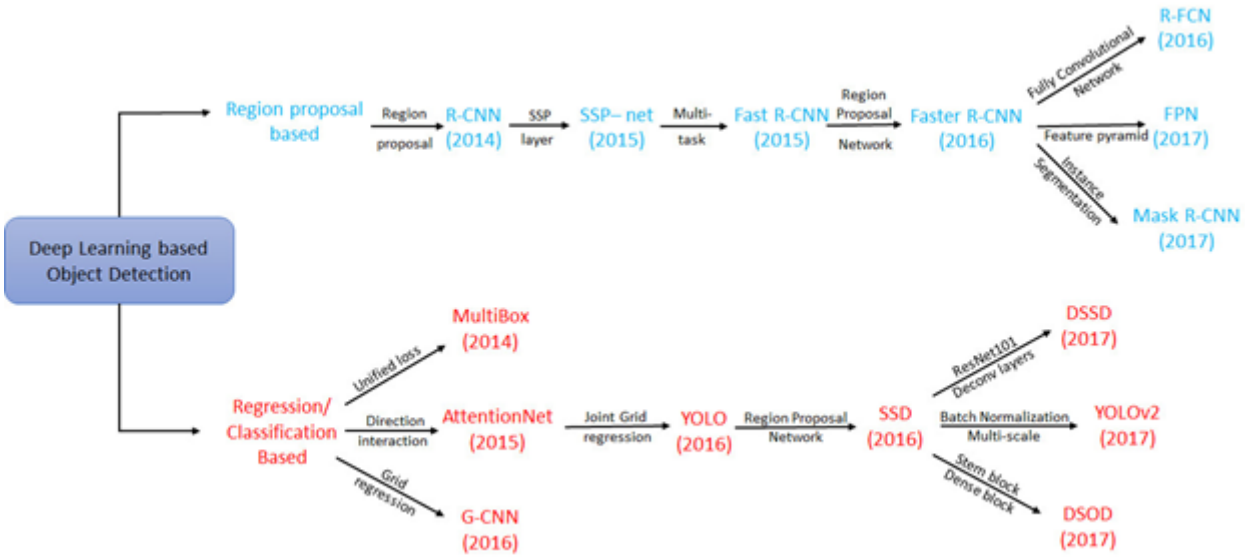
Hình 2.3: Bài toán phát hiện nhiều đối tượng trong ảnh

Một phương pháp tiếp cận sơ khai là chia Hình ảnh thành nhiều vùng quan tâm hoặc sử dụng cửa sổ trượt, sau đó áp dụng CNN để kiểm tra sự hiện diện của đối tượng trên từng vùng. Tuy nhiên, do đối tượng có thể xuất hiện ở bất kỳ vị trí nào với tỷ lệ khung Hình và kích thước vô cùng đa dạng, phương pháp này tạo ra một số lượng vùng xét nghiệm khổng lồ, dẫn đến chi phí tính toán quá cao và không khả thi trong thực tế.

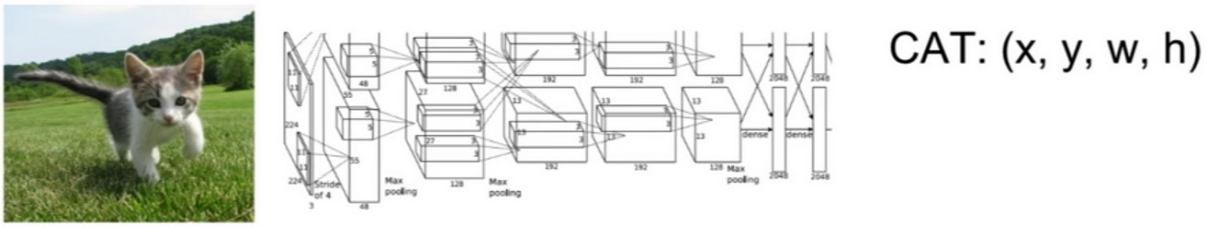
Để vượt qua giới hạn này, các kiến trúc học sâu tiên tiến cho bài toán phát hiện đối tượng đã ra đời và chủ yếu được phân loại thành hai hướng tiếp cận chính: các mô hình dựa trên vùng đề xuất và các mô hình dựa trên hồi quy/phân loại, như được hệ thống hóa trong Hình 2.4.

Phát hiện đơn đối tượng

Phát hiện đơn đối tượng là bài toán nền tảng trong lĩnh vực phát hiện đối tượng (Hình 2.5), với mục tiêu chỉ trích xuất thông tin của một đối tượng duy nhất trong ảnh. Bài toán này là sự kết hợp giữa phân loại và định vị đối tượng. Cụ thể, ngõ ra của mô hình không chỉ cung cấp kết quả phân loại (nhãn của đối tượng) mà còn phải trả về các tọa độ không gian để xác định vị trí chính xác của đối tượng đó



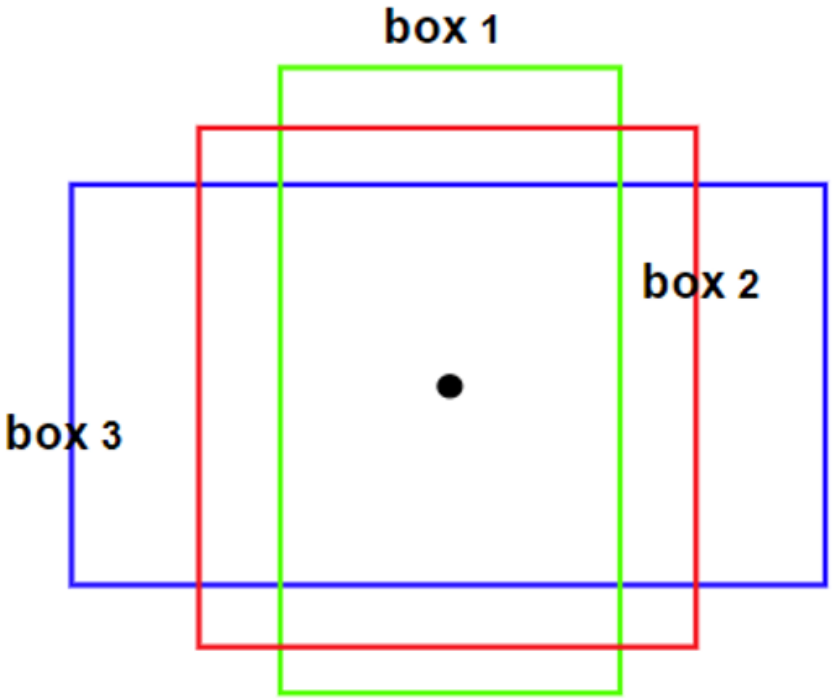
Hình 2.4: Phát hiện đối tượng dựa trên vùng đề xuất và phát hiện đối tượng dựa trên hồi quy/phân loại



Hình 2.5: Bài toán phát hiện đơn đối tượng trong ảnh

Để xác định vị trí, người ta không trực tiếp xác định giá trị tuyệt đối của hộp đối tượng trong ảnh mà thường xác định thông qua các dịch chuyển tương đối của các khung bao (anchor box). Hình 2.6 minh họa một số khung bao trong phát hiện đối tượng. Dựa theo tâm của ảnh, các khung chuẩm thường được định nghĩa trước và các khung thực tế sẽ được dự đoán dựa vào sự thay đổi tương đối của khung bao.

Thay vì dự đoán trực tiếp tọa độ tuyệt đối của khung bao, mô hình học sâu ước lượng độ lệch (offset) giữa khung bao và khung thật (ground truth). Giả sử điểm neo có tâm tại (x_a, y_a) với chiều rộng (w_a) và chiều cao (h_a), còn khung thật có tâm (x^*, y^*) cùng kích thước (w^*, h^*). Các giá trị độ lệch mục tiêu trong quá trình huấn luyện được xác định theo công thức: $(t_x^* = \frac{x^* - x_a}{w_a})$, $(t_y^* = \frac{y^* - y_a}{h_a})$, $(t_w^* = \log(\frac{w^*}{w_a}))$, và $(t_h^* = \log(\frac{h^*}{h_a}))$. Khi mô hình dự đoán được các giá trị (t_x, t_y, t_w, t_h), chúng được sử dụng để tái tạo lại khung bao dự đoán thông qua các công thức ngược: $x = t_x \cdot w_a + x_a$, $y = t_y \cdot h_a + y_a$, $w = w_a \cdot e^{t_w}$, và $h = h_a \cdot e^{t_h}$. Cách biểu diễn này giúp mạng học được mối quan hệ tương đối giữa đối tượng và điểm neo, đồng thời ổn định quá trình huấn luyện bằng cách sử dụng phép biến đổi logarithm cho tỷ lệ kích



Hình 2.6: Khung bao để phát hiện khung bao cụ thể

thước.

2.1.5 Tổng quan các phương pháp phát hiện đối tượng

Trong lĩnh vực thị giác máy tính, các phương pháp phát hiện đối tượng thường được chia thành hai nhóm chính dựa trên kiến trúc xử lý: phương pháp hai giai đoạn (two-stage) và phương pháp một giai đoạn (one-stage).

Các phương pháp hai giai đoạn, với đại diện tiêu biểu là dòng R-CNN (Region-based Convolutional Neural Networks) bao gồm R-CNN, Fast R-CNN và Faster R-CNN, hoạt động dựa trên cơ chế tách biệt. Trước tiên, mô hình tạo ra các vùng đề xuất (Region Proposals) có khả năng chứa đối tượng, sau đó tinh chỉnh và phân loại các vùng này. Mặc dù đạt độ chính xác cao, nhược điểm chí mạng của nhóm này là tốc độ xử lý chậm và yêu cầu tài nguyên tính toán lớn, gây khó khăn cho các ứng dụng giám sát thời gian thực [14].

Ngược lại, các phương pháp một giai đoạn như SSD (Single Shot MultiBox Detector) và các phiên bản YOLO (You Only Look Once) đời đầu xem bài toán phát hiện đối tượng là một bài toán hồi quy trực tiếp. mô hình dự đoán đồng thời vị trí khung bao và xác suất lớp từ Hình ảnh đầu vào trong một lần tính toán duy nhất. Cách tiếp cận này cải thiện đáng kể tốc độ xử lý, tuy nhiên thường gặp thách

thức trong việc cân bằng giữa độ chính xác định vị và khả năng phát hiện các đối tượng kích thước nhỏ [15, 16].

Luận án này tập trung nghiên cứu và ứng dụng hai hướng tiếp cận hiện đại nhất hiện nay nhằm khắc phục các hạn chế trên: kiến trúc YOLOX (đại diện cho dòng one-stage tiên tiến) và Deformable DETR (đại diện cho dòng Transformer).

Tổng quan và nguyên lý hoạt động của mạng YOLO

YOLO (You Only Look Once) là một kiến trúc mạng nơ-ron tích chập (CNN) nổi bật với khả năng phát hiện đối tượng theo thời gian thực nhờ cơ chế xử lý toàn cục bức ảnh trong một lần truyền qua mạng. Ảnh đầu vào được chia thành một lưới kích thước $S \times S$. Tại mỗi ô lưới, mô hình dự đoán B hộp bao cùng với độ tin cậy và C xác suất phân loại lớp.

Do đó, đầu ra của YOLO được biểu diễn dưới dạng một tensor có kích thước như phương trình (2.1).

$$S \times S \times (B \times 5 + C) \tag{2.1}$$

Trong đó, mỗi hộp bao chứa 5 thông tin cơ bản: tọa độ tâm (x, y) tương đối với ô lưới, kích thước (w, h) tương đối với toàn bộ ảnh và điểm độ tin cậy (confidence score) phản ánh mức độ chính xác của dự đoán.

Hàm mục tiêu cho bài toán phát hiện đối tượng

Để tối ưu hóa quá trình học, YOLO sử dụng một hàm mục tiêu tổng hợp từ ba thành phần chính: mất mát định vị (Localization loss) tính toán sai số tọa độ hộp bao, mất mát độ tin cậy (Confidence loss) đánh giá khả năng chứa đối tượng, và mất mát phân loại (Classification loss) cho nhãn đối tượng.

Hàm mục tiêu tổng được thể hiện trong phương trình (2.2).

$$
\begin{aligned}
Loss =& \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} l_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} l_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\widehat{w}_i} \right)^2 + \left(\sqrt{w_i} - \sqrt{\widehat{w}_i} \right)^2 \right] \\
&+ \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} l_{ij}^{obj} \left(C_i - \hat{C}_i \right)^2 + \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} l_{ij}^{noobj} \left(C_i - \hat{C}_i \right)^2 \\
&+ \sum_{i=0}^{s^2} l_i^{obj} \sum_{c \in \text{ classes}} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
\tag{2.2}
$$

Trong biểu thức này, đại lượng l_{ij}^{obj} nhận giá trị 1 nếu hộp bao thứ j tại ô thứ i có chứa đối tượng và ngược lại là l_{ij}^{noobj}. Thành phần l_i^{obj} bằng 1 nếu ô i chứa đối tượng. Các hệ số trọng số λ_{coord} và λ_{noobj} được sử dụng để cân bằng tầm quan trọng giữa dự đoán tọa độ và dự đoán nền, giúp mô hình hội tụ ổn định hơn khi số lượng ô không chứa đối tượng thường áp đảo trong một bức ảnh.

Cải tiến điểm neo (Anchor box) và thuật toán Non-max suppression

Từ phiên bản YOLOv2 trở đi, kiến trúc điểm neo (điểm neo) được áp dụng nhằm thay thế việc dự đoán trực tiếp tọa độ bằng cách dự đoán độ lệch (offset) so với các hộp cơ sở định trước.

Quá trình tính toán tọa độ và kích thước thực tế của hộp bao dự đoán (b_x, b_y, b_w, b_h) và độ tự tin $\sigma(t_0)$ dựa trên các tham số đầu ra của mạng $(t_x, t_y, t_w, t_h, t_0)$ và điểm neo (p_w, p_h) tại ô lưới bị lệch (c_x, c_y) được thực hiện thông qua các phương trình từ (2.3) đến (2.7).

$$b_x = \sigma(t_x) + c_x \tag{2.3}$$

$$b_y = \sigma(t_y) + c_y \tag{2.4}$$

$$b_w = p_w e^{t_w} \tag{2.5}$$

$$b_h = p_h e^{t_h} \tag{2.6}$$

$$Pr_{(object)}, IoU_{(b,object)} = \sigma(t_0) \tag{2.7}$$

Hàm sigmoid σ ở đây giúp giới hạn tọa độ tâm hộp bao nằm gọn trong phạm vi của ô lưới dự đoán, trong khi hàm mũ được dùng để tính toán co giãn kích thước giúp mạng ổn định hơn. Sau khi sinh ra

hàng loạt hộp bao dự đoán, thuật toán Non-max suppression (NMS) được áp dụng để lọc bỏ các dự đoán có độ tin cậy thấp, đồng thời loại bỏ các hộp bao chồng chéo lên nhau (dựa trên chỉ số IoU cao), từ đó chỉ giữ lại một hộp bao duy nhất chính xác nhất cho mỗi đối tượng.

mô hình YOLOF

Ở các biến thể cải tiến sau này như YOLOF, để khắc phục gánh nặng tính toán và bộ nhớ của mạng kim tự tháp đặc trưng (FPN), một bộ giải mã đơn ngõ vào - ngõ ra (SiSo) kết hợp với phép nhân chập giãn nở (Dilated convolution) đã được sử dụng. Phương pháp này cho phép mạng trích xuất các đặc trưng đa tỷ lệ tại nhiều điểm cách xa nhau một cách hiệu quả, kết hợp cùng cơ chế Uniform Matching để giải quyết mất cân bằng mẫu, từ đó đơn giản hóa kiến trúc mạng mà vẫn duy trì được độ chính xác phát hiện đối tượng ở tốc độ cao.

2.1.6 Kiến trúc YOLOX và cơ chế không điểm neo (Anchor-free)

YOLOX [2] là một bước tiến quan trọng trong dòng các mô hình YOLO, loại bỏ sự phụ thuộc vào các điểm neo (điểm neo) được thiết kế thủ công – một yếu tố vốn gây ra nhiều vấn đề về tối ưu hóa và độ phức tạp tính toán trong các phiên bản trước.

Cơ chế Anchor-free và SimOTA

Khác với các phương pháp dựa trên anchor yêu cầu xác định trước kích thước và tỷ lệ khung mẫu (gây khó khăn khi đối tượng có Hình dạng bất thường), YOLOX áp dụng cơ chế anchor-free. Cơ chế này dự đoán trực tiếp bốn giá trị của hộp giới hạn (tọa độ tâm, chiều rộng, chiều cao) tại mỗi vị trí điểm ảnh trên bản đồ đặc trưng.

Để giải quyết vấn đề gán nhãn mẫu dương/âm (label assignment) trong môi trường anchor-free, YOLOX sử dụng chiến lược SimOTA (Simplified Optimal Transport Assignment). SimOTA tự động xác định các mẫu dương tối ưu dựa trên sự cân bằng giữa chi phí phân loại và chi phí định vị, giúp mô hình học tốt hơn các đối tượng bị che khuất hoặc có kích thước thay đổi liên tục trên biển.

Kiến trúc đầu tách

Một cải tiến đáng chú ý khác của YOLOX là việc sử dụng kiến trúc đầu tách. Trong các phiên bản YOLO cũ, tác vụ phân loại (classification) và hồi quy vị trí (localization) được thực hiện trên cùng một nhánh tích chập, dẫn đến xung đột tối ưu hóa. YOLOX tách biệt hai tác vụ này thành hai nhánh

song song sau lớp trích xuất đặc trưng. Thực nghiệm cho thấy kiến trúc này giúp mô hình hội tụ nhanh hơn và cải thiện đáng kể độ chính xác mAP (mean Average Precision).

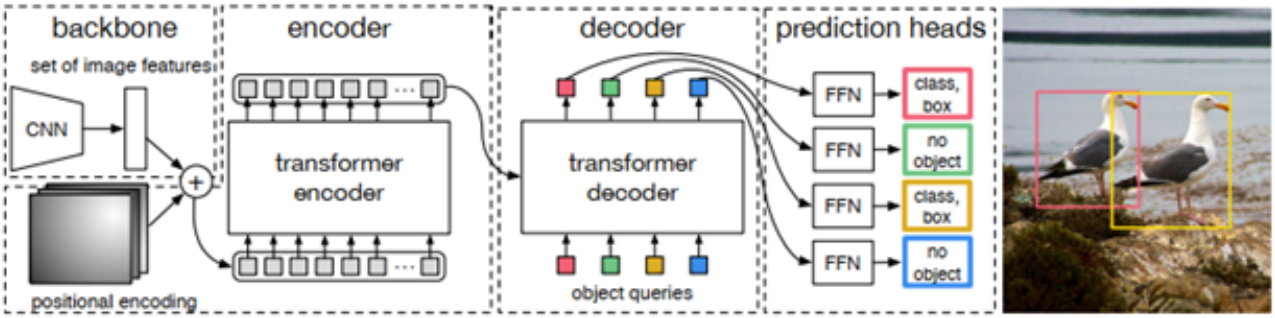
2.1.7 mô hình phát hiện đối tượng dựa trên Transformer

Cùng với sự thành công của mô hình transformer trong xử lý ngôn ngữ tự nhiên, nhiều nhà nghiên cứu cũng áp dụng mô hình transformer vào bài toán phát hiện đối tượng. Một trong những phương pháp phát hiện đối tượng dự trên transformer nổi tiếng nhất là DETR [17]. mô hình này bao gồm một bộ trích xuất đặc trưng, một bộ mã hóa (encoder), và một bộ giải mã (decoder), như được minh họa trong Hình 2.7.

Bộ trích xuất đặc trưng là một mạng CNN backbone có nhiệm vụ trích xuất thông tin mức cao từ ảnh; mã hóa vị trí Hình sin 2D cũng được sử dụng để mã hóa thông tin vị trí cho từng pixel. Đặc trưng Hình ảnh và đặc trưng vị trí được ghép nối và đưa vào bộ mã hóa dựa trên transformer.

Bộ mã hóa bao gồm nhiều lớp self-attention đa đầu (multi-head) xếp chồng lên nhau. Đặc trưng từ bộ mã hóa sau đó được truyền đến bộ giải mã dựa trên transformer. Bộ giải mã cũng nhận một tập các truy vấn đối tượng (object queries) có thể học được làm đầu vào. Những truy vấn này đóng vai trò như các latent biểu diễn các vị trí tiềm năng trên ảnh. Với mỗi truy vấn, bộ giải mã sẽ sử dụng đặc trưng từ bộ mã hóa để dự đoán xem có đối tượng nào tại vị trí đó hay không.

Trong bộ trích xuất đặc trưng, với đầu vào là một ảnh có kích thước tensor $H_0 \times W_0 \times 3$, backbone CNN sẽ tạo ra bản đồ đặc trưng có kích thước $H \times W \times C$. Theo [18], ta có $C = 2048$, $H = H_0/32$, và $W = W_0/32$. Một lớp tích chập 1x1 được sử dụng để giảm số kênh đặc trưng từ C xuống d $(d < C)$, tạo ra bản đồ đặc trưng mới có kích thước $H \times W \times d$. Vì bộ mã hóa dựa trên transformer yêu cầu đầu vào 2D, bản đồ đặc trưng được biến đổi thành ma trận kích thước $d \times HW$. Các hàng của ma trận này được gọi là *tokens*, mỗi token là một vector d chiều.



Hình 2.7: Kiến trúc DETR

Bộ mã hóa bao gồm nhiều lớp self-attention đa đầu xếp chồng. Mỗi khối multi-head self-attention gồm nhiều mô-đun self-attention. Trong một mô-đun self-attention đa đầu, các đặc trưng được trích xuất từ các mô-đun attention riêng lẻ sẽ được ghép nối và chiếu tuyến tính ra đầu ra. Ba phép chiếu tuyến tính độc lập được sử dụng để trích xuất bộ ba $(key, value, query)$ trong một mô-đun self-attention. Độ tương đồng giữa *key* và *query* đóng vai trò là hệ số attention cho *value*, giúp hợp nhất *value* thành đặc trưng attention tổng hợp.

Bộ giải mã nhận đặc trưng mới từ bộ mã hóa và các truy vấn đối tượng có thể học được. Các đặc trưng từ bộ mã hóa thể hiện thông tin Hình ảnh tại mọi vị trí, trong khi các truy vấn thể hiện mã hóa vị trí có thể học, đặt câu hỏi liệu có đối tượng tại vị trí cụ thể nào không. Bộ giải mã sử dụng thông tin Hình ảnh để xác định xem vị trí được mã hóa có chứa lớp nào hay không. Các truy vấn này được học từ dữ liệu huấn luyện. Kiến trúc của các mô-đun bộ mã hóa và bộ giải mã được minh họa trong Hình 2.7.

Đầu ra của mô-đun giải mã được đưa vào một mạng nơ-ron lan truyền tiến (feedforward neural network) để dự đoán vị trí và loại của đối tượng. Các đầu ra này tương ứng với từng truy vấn đối tượng được học trong khối giải mã. Nếu một truy vấn đối tượng không mã hóa bất kỳ đối tượng nào, đầu ra của nhánh phân loại sẽ là "No Object".

Mạng DETR tạo ra một tập N kết quả phát hiện tương ứng với N truy vấn đối tượng. Mỗi kết quả là một bộ $(class, box)$ đại diện cho duy nhất một đối tượng mà không bị trùng lặp. Do đó, cần một quá trình ghép cặp (matching) để ánh xạ mỗi dự đoán với một nhãn thật. Vì số lượng đối tượng thật trong ảnh nhỏ hơn số lượng truy vấn N, nên một tập $\varnothing$ gồm các vùng ảnh được cắt ngẫu nhiên từ nền được thêm vào làm nhãn thật bổ sung.

Do đó, các "đối tượng nền" với vị trí ngẫu nhiên và nhãn "No Object" giúp cân bằng số lượng nhãn thật và số lượng dự đoán. Ký hiệu σ là lời giải ghép cặp; ghép cặp tối ưu $\hat{\sigma}$ được xác định thông qua quá trình tối ưu như trong phương trình 2.8.

$$\hat{\sigma} = \underset{\sigma \in \Xi_N}{\operatorname{argmin}} \sum_{i}^{N} L_{\text{match}} \left(y_i, \hat{y}_{\sigma(i)}\right) \tag{2.8}$$

Trong đó y_i là nhãn thật của một hộp bao gồm nhãn lớp và nhãn hộp (c_i, b_i); và $\hat{y}_i$ là kết quả dự đoán. Quá trình gán tối ưu này được giải bằng thuật toán Hungarian [19]. Lưu ý rằng chi phí này không được tính riêng lẻ cho từng đối tượng, mà cho toàn bộ tập N đối tượng của một ảnh.

Một mô hình tốt có thể dự đoán chính xác các đối tượng và hộp bao với độ chồng lấp cao hơn. Do đó,

hàm mục tiêu $L_m\left(y_i, \hat{y}_{\sigma(i)}\right)$ cần bao gồm các tiêu chí này như trong phương trình 2.9.

$$L_m\left(y_i, \hat{y}_{\sigma(i)}\right) = -1_{\{c_i \neq \varnothing\}} \log \hat{p}_{\sigma(i)}\left(c_i\right) + 1_{\{c_i \neq \varnothing\}} L_{box}\left(b_i, \hat{b}_{\sigma(i)}\right) \tag{2.9}$$

Trong đó:

- $L_m(y_i, \hat{y}_{\sigma(i)})$: là hàm chi phí gán (matching cost) đánh giá mức độ sai khác giữa đối tượng thực tế thứ i và kết quả dự đoán tương ứng được gán bởi phép hoán vị σ.
- $y_i = (c_i, b_i)$: là tập hợp thông tin nhãn thực tế (ground truth) của đối tượng thứ i, bao gồm nhãn lớp c_i và tọa độ hộp bao b_i.
- $\hat{y}_{\sigma(i)}$: là kết quả dự đoán của mô hình được gán cho đối tượng thực tế thứ i thông qua thuật toán gán tối ưu Hungarian.
- c_i: là nhãn lớp thực tế của đối tượng; $\varnothing$ ký hiệu cho lớp nền (background), đại diện cho trường hợp không có đối tượng.
- $1_{\{c_i \neq \varnothing\}}$: là hàm chỉ thị (indicator function), nhận giá trị bằng 1 nếu mẫu dữ liệu thực tế là một đối tượng và bằng 0 nếu là lớp nền.
- $\hat{p}_{\sigma(i)}(c_i)$: là xác suất dự đoán của mô hình cho lớp c_i tại vị trí chỉ số dự đoán $\sigma(i)$.
- b_i và $\hat{b}_{\sigma(i)}$: lần lượt là tọa độ không gian (vị trí tâm, chiều rộng, chiều cao) của hộp bao thực tế và hộp bao dự đoán.
- $L_{box}(b_i, \hat{b}_{\sigma(i)})$: là hàm mất mát không gian hộp bao, thường được cấu thành từ sự kết hợp tuyến tính của sai số chuẩn hóa L_1 và chỉ số giao của phần hợp tổng quát (Generalized IoU - GIoU).

Thành phần đầu tiên của Phương trình 2.9 hướng dẫn mô hình dự đoán chính xác loại của đối tượng. Thành phần thứ hai giúp dự đoán hộp bao chính xác hơn. Ở đây $c_i \neq \varnothing$ ám chỉ việc phát hiện các đối tượng thật (không phải đối tượng giả). Vì mỗi đối tượng thật chỉ được phát hiện một lần, phương trình 2.9 được áp dụng cho tất cả các đối tượng thật trong ảnh.

Ghép cặp tối ưu $\hat{\sigma}$ được ước lượng trên tập bao gồm cả các đối tượng thật và đối tượng giả (được định nghĩa bởi $\varnothing$). Vì hộp của các đối tượng giả không có ý nghĩa, nên thành phần box loss chỉ hợp lệ khi hộp không thuộc tập $\varnothing$. Ngược lại, các đối tượng giả nên được dự đoán chính xác là "No Object". Do đó, thành phần phân loại cần bao gồm cả tập $\varnothing$. Dựa trên quan sát này, hàm mất mát Hungary (L_H) [19] trong Phương trình 2.10 được sử dụng để huấn luyện mô hình.

$$L_H(y,\hat{y}) = \sum_{i=1}^{N} \left[-\log \hat{p}_{\sigma(i)}(c_i) + 1_{\{c_i \neq \varnothing\}} L_{\text{box}}\left(b_i, \hat{b}_{\sigma(i)}\right) \right], \tag{2.10}$$

Trong đó:

- L_m, L_H: Chi phí gán và tổng mất mát Hungary.
- $y_i, \hat{y}_{\sigma(i)}$: Nhãn thực và kết quả dự đoán được gán cặp tương ứng.
- c_i, b_i: Nhãn lớp và tọa độ hộp bao thực tế; $\varnothing$ là lớp nền.
- $1_{\{c_i \neq \varnothing\}}$: Hàm chỉ thị, bằng 1 nếu là đối tượng thật, ngược lại bằng 0.
- $\hat{p}_{\sigma(i)}(c_i)$: Xác suất mô hình dự đoán đúng lớp c_i.
- L_{box}: Mất mát hộp bao (tổng hợp từ L_1-norm và $GIoU$).
- N: Tổng số lượng truy vấn đối tượng trong một ảnh.

Tỷ lệ ảnh hưởng (influence scaling) là yếu tố quan trọng trong việc ước lượng hộp bao. Ví dụ, hộp bao của một đối tượng lớn có thể có chiều rộng 0.2, trong khi hộp bao của một đối tượng nhỏ chỉ 0.02. Nếu sử dụng khoảng cách Euclide thông thường để đo lường mất mát hộp bao, mô hình có thể bị thiên lệch quá nhiều về các đối tượng lớn và bỏ qua các đối tượng nhỏ. Do đó, hàm mất mát Generalized Intersection over Union (GIOU) [20] được giới thiệu để tính toán mất mát hộp bao kết hợp với hàm mất mát L1-norm [21]. Ký hiệu $\lambda_{iou}, \lambda_{L1} \in R$ là các siêu tham số điều chỉnh quá trình học cho GIOU loss và L1-norm loss; công thức của hàm mất mát biên được thể hiện trong Phương trình 2.11.

$$L_{\text{box}}\left(b_i, \hat{b}_{\sigma(i)}\right) = \lambda_{\text{iou}} L_{\text{iou}}\left(b_i, \hat{b}_{\sigma(i)}\right) + \lambda_{L1} \left\| b_i - \hat{b}_{\sigma(i)} \right\|_1 \tag{2.11}$$

2.1.8 Các chỉ số đánh giá cho bài toán phát hiện đối tượng đối tượng

Đối với mỗi mô hình phát hiện đối tượng, sau khi huấn luyện, cần có những tham số để đánh giá độ chính xác của chúng. Hiện nay, mAP (mean average precision) là một chỉ số được sử dụng phổ biến cho bài toán phát hiện đối tượng. Để tìm hiểu về mAP, trước tiên ta phải nắm được những khái niệm như độ chồng lần của hai đối tượng (IoU), khả năng dự đoán mẫu dương tính (precision), và độ nhạy (recall).

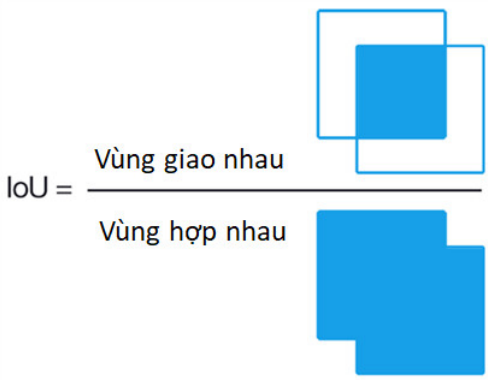
Trong quá trình huấn luyện mô hình phát hiện đối tượng, dữ liệu đầu vào bao gồm hai thành phần: Hình ảnh và các hộp bao được gắn nhãn cho từng đối tượng trong Hình. Các đường bao gắn nhãn này

thường được tạo ra bởi con người thông qua các công cụ gắn nhãn. Sau khi huấn luyện, mô hình sẽ dự đoán các đường bao cho những đối tượng được phát hiện trong ảnh.

Độ chồng lấn (IoU – Intersection over Union) là chỉ số dùng để so sánh mức độ trùng khớp giữa đường bao dự đoán và đường bao gắn nhãn. IoU được tính bằng tỉ lệ giữa diện tích phần giao nhau và diện tích phần hợp của hai đường bao, như minh họa ở Hình 2.8.

Dựa trên giá trị IoU, việc đánh giá mô hình được xác định theo các tiêu chí sau:

- Nếu IoU lớn hơn ngưỡng cho trước: đối tượng được xem là nhận dạng đúng ⇒ True Positive (TP).
- Nếu IoU nhỏ hơn ngưỡng: đối tượng bị nhận dạng sai ⇒ False Positive (FP).
- Nếu một đối tượng trong ảnh không được phát hiện: ⇒ False Negative (FN).



Hình 2.8: Intersection over Union

Khả năng dự đoán mẫu dương tính (precision) và độ nhạy (recall)

Khả năng dự đoán mẫu dương tính (precision) được sử dụng để đánh giá độ tin cậy của mô hình, tức là tỉ lệ phần trăm các dự đoán dương tính mà mô hình đưa ra là chính xác. Precision được xác định theo phương trình 2.12:

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{2.12}$$

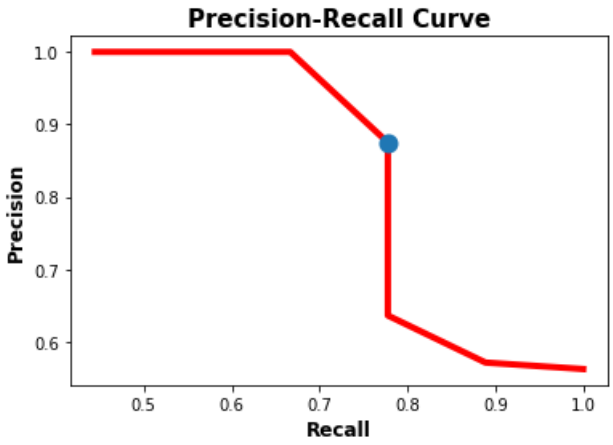
Độ nhạy (recall) phản ánh khả năng của mô hình trong việc phát hiện đầy đủ các đối tượng thực tế (ground truth), tức là tỉ lệ phần trăm các mẫu dương tính thực sự được mô hình nhận dạng. Recall được tính theo phương trình 2.13:

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2.13}$$

trong đó:

- *True Positive*: Số lượng đối tượng thực tế được mô hình dự đoán chính xác cả về vị trí và nhãn lớp.
- *False Negative*: Số lượng đối tượng thực tế tồn tại trong ảnh nhưng mô hình không phát hiện được hoặc dự đoán sai.

Mối quan hệ giữa precision và recall thường được biểu diễn thông qua đường cong Precision–Recall (Hình 2.9). Đường cong này thể hiện sự thay đổi của hai chỉ số ứng với từng mức ngưỡng IoU khác nhau, giúp ta lựa chọn ngưỡng phù hợp sao cho cả precision và recall đều đạt giá trị cao.



Hình 2.9: Ví dụ minh họa đường cong Precision–Recall [10]

Tuy nhiên, việc lựa chọn ngưỡng tối ưu trực tiếp trên đồ thị chỉ thuận lợi khi đường cong không quá phức tạp. Một phương pháp phổ biến và hiệu quả hơn là sử dụng chỉ số F1, được định nghĩa như phương trình 2.14:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2.14}$$

Trong đó:

- $F1$: Giá trị trung bình điều hòa dùng để đánh giá sự cân bằng giữa độ chính xác và độ nhạy của mô hình.
- *precision*: Tỉ lệ số lượng đối tượng dự đoán đúng trên tổng số các đối tượng mà mô hình đã phát

hiện.

- *recall*: Tỉ lệ số lượng đối tượng dự đoán đúng trên tổng số các đối tượng thực tế tồn tại trong dữ liệu.

Chỉ số F1 đo lường sự cân bằng giữa độ chính xác (precision) và tỷ lệ thu hồi (recall). Giá trị F1 cao cho thấy cả hai chỉ số đều ở mức tốt, trong khi giá trị F1 thấp phản ánh sự mất cân bằng, tức một trong hai chỉ số bị giảm đáng kể so với chỉ số còn lại.

Ví dụ sau minh hoạ cách tính giá trị F1 score: Giả sử trong một tập kiểm thử, mô hình phát hiện được:

- **True Positive (TP) = 80**: số đối tượng được nhận dạng đúng,
- **False Positive (FP) = 20**: số đối tượng mô hình dự đoán sai,
- **False Negative (FN) = 10**: số đối tượng thực tế có nhưng mô hình bỏ sót.

Khi đó, các chỉ số được tính như sau:

$$precision = \frac{TP}{TP+FP} = \frac{80}{80+20} = 0{,}8 \tag{2.15}$$

$$recall = \frac{TP}{TP+FN} = \frac{80}{80+10} \approx 0{,}889 \tag{2.16}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = 2 \cdot \frac{0{,}8 \cdot 0{,}889}{0{,}8 + 0{,}889} \approx 0{,}842 \tag{2.17}$$

Kết quả cho thấy: *precision* đạt 0,8 (tức 80% đối tượng mà mô hình dự đoán dương tính là chính xác), *recall* đạt khoảng 0,889 (mô hình phát hiện được gần 89% đối tượng thực tế), và *F1-score* là 0,842, phản ánh sự cân bằng tương đối tốt giữa *precision* và *recall*.

Bên cạnh các độ đo trên, tổng số lượng đối tượng (hoặc số lượng hộp bao - bounding boxes) mà mô hình phát hiện được trong tập dữ liệu thử nghiệm cũng là một thông số quan trọng. Trong nghiên cứu này, số lượng phát hiện được ký hiệu là $n_{detected}$

AP và mAP

Giá trị Average Precision (AP) đại diện cho diện tích nằm phía dưới đường cong Precision–Recall. Cần lưu ý rằng mặc dù đây là một dạng diện tích dưới đường cong, nhưng đại lượng này được định

nghĩa chuyên biệt để phản ánh hiệu quả dự đoán trung bình của mô hình và phân biệt rõ với các loại diện tích dưới đường cong khác như ROC-AUC. Chỉ số này đóng vai trò là giá trị tóm tắt giúp đánh giá năng lực của hệ thống trên toàn bộ các ngưỡng thay đổi của Recall. Công thức tính toán được xác định theo phương trình 2.18:

$$AP = \sum_{k=0}^{k=n-1} \left[recall_k - recall_{k+1}\right] \cdot precision_k \tag{2.18}$$

Trong đó:

- n là số lượng ngưỡng (thresholds),
- k là chỉ số của cặp điểm $(recall_k, precision_k)$,
- $recall_k, precision_k$ lần lượt là độ nhạy và độ chính xác tại ngưỡng k^{th}.

2.2 Phân cụm bằng kỹ thuật học sâu

2.2.1 Tổng quan về phân cụm

Phân cụm là một bài toán cơ bản trong học máy và thường đóng vai trò là bước tiền xử lý quan trọng trong nhiều tác vụ khai thác dữ liệu. Mục đích chính của phân cụm là tác các mẫu dữ liệu thành các nhóm sao cho các mẫu giống nhau thuộc về cùng một cụm trong khi các mẫu khác nhau thuộc về các cụm khác nhau. Các cụm mẫu cung cấp đặc tính toàn cục của phân phối dữ liệu, có thể mang lại các kiến thức chuyên sâu trên toàn bộ tập dữ liệu, chẳng hạn như phát hiện sự bất thường [22] và học tập các đặc trưng phân biệt [23] v.v.

Ngày nay, các kỹ thuật phân cụm thường được chia thành hai nhóm. Các phương pháp phân cụm nông [22, 24] (shallow clustering) đã đạt được sự thành công lớn, nhưng chúng giả định rằng các mẫu dữ liệu đã được biểu diễn trong không gian đặc trưng với một phân phối đủ tốt. Với sự phát triển nhanh chóng của internet và dịch vụ web trong thập kỷ qua, cộng đồng nghiên cứu đang thể hiện sự quan tâm ngày càng tăng về việc tìm hiểu các mô hình học máy mới có khả năng xử lý dữ liệu không có đặc trưng rõ ràng, chẳng hạn như Hình ảnh và dữ liệu có số chiều cao lên đến hàng nghìn đặc trưng, v.v. Do đó, các phương pháp phân cụm nông không còn áp dụng trực tiếp cho việc xử lý dữ liệu như vậy.

Những năm gần đây, kỹ thuật học sâu đã đạt được sự thành công đáng kể trong việc học các đặc trưng

mô tả dữ liệu. Tuy nhiên, việc áp dụng kỹ thuật học sâu chưa được thảo luận nhiều trong bài toán phân cụm. Nguyên nhân là do sự thiếu hụt các kiến thức tiền nghiệm về đặc điểm của phân cụm tốt cho một loại dữ liệu cụ thể.

Để giải quyết các thách thức này, kỹ thuật phân cụm dựa trên học sâu (Deep Clustering) đã được phát triển. Mục tiêu của phương pháp này là tối ưu hóa đồng thời cả hai nhiệm vụ: rút trích các đặc trưng và phân cụm các mẫu. Cụ thể, các phương pháp phân cụm dựa trên học sâu tập trung vào việc nghiên cứu các thách thức sau đây:

(1) Làm thế nào để học các đặc trưng tốt hơn, nhờ đó có thể mang lại hiệu suất phân cụm tốt hơn?

(2) Làm thế nào để thực hiện phân cụm và rút trích đặc trưng một cách hiệu quả trong một khuôn khổ thống nhất?

(3) Làm thế nào để phá vỡ rào cản giữa phân cụm và rút trích đặc trưng, cho phép chúng tương tác và tối ưu hóa lẫn nhau trong suốt quá trình huấn luyện?

Để giải quyết các thách thức trên, nhiều phương pháp phân cụm dựa trên học sâu đã được đề xuất với các kiến trúc khác nhau và nhiều biến thể dữ liệu. Mặc dù vậy, mỗi biến thể đều xoay quanh các câu hỏi lớn bao gồm:

- Quy trình rút trích đặc trưng.
- Quy trình phân cụm.
- Sự tương quan giữa hai quy trình này trong suốt quá trình huấn luyện.

2.2.2 Phân cụm dữ liệu truyền thống

Bài toán phân cụm dữ liệu thông thường sẽ bao gồm các câu hỏi kỹ thuật sau:

- Các đặc trưng nào sẽ được sử dụng? Các đặc trưng có thể là tín hiệu thô từ cảm biến hoặc các đặc trưng do chuyên gia thiết kế. Bên cạnh đó, các bước tiền xử lý như chuẩn hoá dữ liệu hoặc giảm chiều dữ liệu cũng sẽ được dùng để tăng độ hiệu quả của thuật toán. Đây là bước cực kỳ quan trọng vì nếu các đặc trưng tốt được rút trích, thì việc phân cụm sẽ rất dễ dàng. Một số phương án rút trích đặc trưng tiêu biểu cho các tín hiệu trong miền thời gian bao gồm: FFT [25], DCT [26], hay DWT [27]. Phương án chuẩn hoá dữ liệu thường là phương án normal scaler [28]. Phương án giảm chiều dữ liệu thường là PCA [29].

- Cách tính khoảng các giữa các mẫu. Ý tưởng chung cho mọi thuật toán phân cụm là khoảng cách từ một điểm đến tâm của nhóm là gần nhất. Tuy nhiên khoảng cách có thể là khoảng cách Euclidean [30] hoặc khoảng cách Manhattan [31]. Thuật toán nổi tiếng k-means [32] dựa vào khoảng cách Euclidean, trong khi các thuật toán phi tuyến như phân cụm Spectral [33] dựa vào khoảng cách Manhattan.
- Cách xác định số lượng các cụm. Thông thường, các thuật toán phân cụm phải tự xác định số cụm trước khi tiến hành phân cụm và phải khởi tạo đại diện của mỗi nhóm. Tuy nhiên một số khác có thể tự tìm số nhóm thông qua sự tương tác tự thân của các điểm dữ liệu như thuật toán Affinity Propagation [34].

2.2.3 Phân cụm dữ liệu dựa trên học sâu

Ngày nay, sự phát triển của kỹ thuật học sâu đã trở thành xu hướng trong mọi ứng dụng của học máy và kỹ thuật phân cụm cũng không ngoại lệ. Một mô hình phân cụm dựa trên học sâu sẽ bao gồm ba thành tố chính. Thứ nhất, làm thế nào để rút trích đặc trưng. Mục đích của bước này là chiếu các mẫu sang một miền đặc trưng có thể hỗ trợ tốt hơn cho việc phân cụm. Thứ hai, làm thế nào để phân cụm. Mục đích của khối này là chiếu thông tin từ miền đặc trưng sang các nhóm. Thứ ba, hai khối rút trích đặc trưng và phân cụm sẽ tương tác với nhau như thế nào.

Để rút trích đặc trưng, một số chiến lược sau sẽ được áp dụng như kỹ thuật học các đặc trưng dựa trên Auto-Encoder [35], kỹ thuật tạo sinh dữ liệu [36], kỹ thuật giả lập mẫu [37], kỹ thuật hướng tâm [38]. Kỹ thuật Auto-Encoder [35] là một dạng của giảm chiều dữ liệu, theo đó dữ liệu được giảm chiều có thể được tái tạo thành dữ liệu nguyên gốc. Đây có thể được coi là một biến thể phi tuyến của các thuật toán cơ bản như PCA. Kỹ thuật tạo sinh dữ liệu [36] có cấu trúc giống với kỹ thuật Auto-Encoder [35] nhưng hướng tới việc học các đặc trưng sao cho chỉ cần lấy mẫu một biến trong miền đặc trưng thì bộ giải mã của mạng sẽ có thể tái tạo lại một mẫu mới có ngữ nghĩa chứ không cần rút trích đặc trưng từ một mẫu thật. Kỹ thuật giả lập mẫu [37] sẽ tạo ra một mẫu mới giả lập từ mẫu gốc; bộ rút trích đặc trưng sẽ phải rút trích các đặc trưng giống nhau cho cả mẫu gốc và mẫu giả lập. Kỹ thuật hướng tâm [38] dựa trên ý tưởng rằng đặc trưng của các mẫu trong cùng một nhóm càng gần hoặc càng giống với tâm đại diện của nhóm thì càng tốt. Do đó, nó sẽ dựa trên tâm cũ để gán nhãn giả cho các mẫu, và sử dùng thuật toán phân loại để huấn luyện lại bộ rút trích đặc trưng.

Để phân cụm dữ liệu, các chiến lược cơ bản sau có thể được kể tới. Chiến lược dựa vào làm khớp

mối liên hệ [39], chiến lược dựa trên các nhãn giả [38], chiến lược tự huấn luyện [40]. Chiến lược làm khớp [39] cho rằng các mẫu có đặc trưng giống nhau sẽ thuộc về các lớp giống nhau. Mặc dù chiến lược này dễ thực hiện nhưng việc tính toán khoảng cách giữa tất cả các cặp điểm sẽ không khả thi trong thực tế. Một số biến thể khác sẽ chỉ tính toán sự tương quan giữa các mẫu lân cận. Chiến lược dựa trên các nhãn giả [38] sử dụng một tâm cho trước để gán nhãn giả cho tất cả các mâũ, sau đó bộ rút trích đặc trưng sẽ được học dựa trên các nhãn giả đó. Nhược điểm chính của phương pháp này là đòi hỏi mô hình rút trích đặc trưng phải đủ tốt; nếu không các nhãn giả sinh ra sẽ bị nhiễu và phá hỏng bộ rút trích đặc trưng. Chiến lược tự huấn luyện [40] đặt ra giả thiết về một phân phối kỳ vọng của dữ liệu sau khi phân cụm; sau đó quá trình huấn luyện sẽ hướng phân phối của kết quả phân cụm vào giống với kết quả phân phối kỳ vòng.

Để kết hợp khối rút trích đặc trưng và khối phân cụm, một số chiến lược sau có thể được áp dụng. Chiến lược nhiều bước [39] chia bước rút trích đặc trưng và bước phân cụm thành hai bước riêng biệt. Ưu điểm của phương pháp này là đơn giản, dễ thực hiện. Nhược điểm là quá trình phân cụm không thể hỗ trợ cho quá tình rút trích đặc trưng nhằm tạo ra các đặc trưng tốt hơn. Chiến lược lặp [38] là một quá trình lặp đi lặp lại hai bước rút trích đặc trưng là phân cụm. Tuy hai bước này sử dụng chung một hàm mục tiêu nhưng quá trình cập nhập trọng số cho mỗi bước được tiến hành một cách tuần tự. Trong phương pháp kinh điển DeepCluster [38], quá trình cập nhập bộ rút trích đặc trưng được thực hiện thông qua các nhãn giả, còn quá trình phân cụm được thực hiện thông qua quá trình cập nhập tâm đại diện của mỗi nhóm. Cuối cùng, chiến lược tích hợp [40] sẽ cho phép cả khối rút trích đặc trưng và khối phân cụm được cập nhập đồng thời. Ưu điểm của chiến lược lặp và chiến lược tích hợp là quá trình phân cụm có thể hỗ trợ thêm cho bộ rút trích đặc trưng. Tuy nhiên nhược điểm của các phương pháp phức tạp này là cần có bộ dữ liệu đủ tốt hoặc các bộ rút trích đặc trưng được huấn luyện bởi một tập dữ liệu rất lớn.

2.2.4 Các chỉ số đánh giá

Các thuật toán phân cụm được đánh giá bằng các chỉ số như thông tin tương hỗ (mutual information - MI) và các biến thể của nó, chỉ số tính đầy đủ (completeness score), chỉ số Fowlkes-Mallows, và chỉ số Rand Index.

Thông tin tương hỗ và các biến thể

Thông tin Tương hỗ là một thước đo lượng thông tin mà một biến ngẫu nhiên chứa về một biến ngẫu nhiên khác. Nói một cách đơn giản, nó định lượng mức độ phụ thuộc giữa hai biến. Trong bài toán

phân cụm, MI có thể được sử dụng để đánh giá mức độ tương đồng giữa kết quả phân cụm và nhãn gốc (ground truth). Điểm MI càng cao thì mức độ đồng thuận giữa kết quả phân cụm và nhãn gốc càng tốt. Điểm số bằng 0 cho thấy không có sự tương đồng nào ngoài mức ngẫu nhiên, và giá trị âm có thể xảy ra nếu kết quả phân cụm tệ hơn so với phân cụm ngẫu nhiên. Công thức toán học của MI giữa hai biến ngẫu nhiên X và Y được định nghĩa trong phương trình 2.19.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2.19}$$

Trong đó:

- $I(X;Y)$ là thông tin tương hỗ giữa hai biến X và Y.
- $p(x,y)$ là phân phối xác suất đồng thời của X và Y.
- $p(x)$ và $p(y)$ là phân phối xác suất biên của X và Y.

MI có xu hướng ưu tiên các thuật toán phân cụm sinh ra nhiều cụm. Do đó, Thông tin tương hỗ điều chỉnh (Adjusted Mutual Information - AMI) và Thông tin tương hỗ chuẩn hoá (Normalized Mutual Information - NMI) thường được sử dụng để hiệu chỉnh theo entropy của nhãn gốc và cụm phân loại, nhằm tránh sự thiên lệch.

$$AMI(X,Y) = \frac{I(X;Y) - E[I(X;Y)]}{\max[H(X),H(Y)] - E[I(X;Y)]} \tag{2.20}$$

$$NMI(X,Y) = \frac{2 \times I(X,Y)}{H(X) + H(Y)} \tag{2.21}$$

Trong đó:

- $H(X)$ và $H(Y)$ là entropy của kết quả phân cụm X và Y.
- $I(X,Y)$ là thông tin tương hỗ giữa hai kết quả phân cụm X và Y.
- $E[I(X,Y)]$ là giá trị kỳ vọng của MI giữa hai phân cụm ngẫu nhiên, được định nghĩa trong Phương trình 2.22.

$$E[I(X,Y)] = \sum_{i=1}^{k} \sum_{j=1}^{l} p(x_i,y_j) \log \left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)} \right) \tag{2.22}$$

Chỉ số AMI hiệu chỉnh điểm số MI theo kỳ vọng ngẫu nhiên, với giá trị dao động từ -1 đến 1. Giá trị

bằng 1 biểu thị sự trùng khớp hoàn hảo, giá trị bằng 0 biểu thị sự trùng khớp ngẫu nhiên, và giá trị âm thể hiện phân cụm còn tệ hơn so với ngẫu nhiên.

Trong khi đó, NMI chuẩn hoá MI bằng cách chia cho trung bình cộng entropy của hai phân cụm, cho giá trị trong khoảng từ 0 (không có sự đồng thuận ngoài ngẫu nhiên) đến 1 (đồng thuận hoàn hảo).

Chỉ số tính đầy đủ

Chỉ số tính đầy đủ ($CompletenessScore - CS$) đánh giá mức độ mà tất cả các điểm dữ liệu thuộc cùng một lớp thực (ground truth) được thuật toán phân cụm gán vào cùng một cụm. Công thức được định nghĩa như sau:

$$CS = \frac{\sum_i \max_j \left|C_i \cap K_j\right|}{\sum_i |C_i|} \tag{2.23}$$

Trong đó:

- C_i là tập dữ liệu thuộc lớp thực i.
- K_j là tập dữ liệu thuộc cụm j.
- $|C_i \cap K_j|$ là số lượng điểm dữ liệu vừa thuộc lớp thực i, vừa thuộc cụm j.
- $|C_i|$ là tổng số điểm dữ liệu trong lớp thực i.

Giá trị $CS = 1{,}0$ biểu thị phân cụm hoàn hảo, trong khi giá trị thấp hơn cho thấy các điểm của cùng một lớp thực bị phân tán vào nhiều cụm khác nhau.

Chỉ số Fowlkes-Mallows (FMI):

Chỉ số Fowlkes-Mallows được sử dụng để đo mức độ tương đồng giữa hai kết quả phân cụm (thường là kết quả dự đoán và nhãn gốc). Nó được định nghĩa là trung bình Hình học giữa precision và recall theo cặp điểm dữ liệu:

$$FMI = \frac{TP}{\sqrt{(TP+FP)(TP+FN)}} \tag{2.24}$$

Trong đó:

- TP (True Positives): số cặp điểm thuộc cùng một cụm trong cả kết quả phân cụm dự đoán và nhãn gốc.
- FP (False Positives): số cặp điểm cùng cụm trong kết quả dự đoán nhưng khác cụm trong nhãn

gốc.

- FN (False Negatives): số cặp điểm cùng cụm trong nhãn gốc nhưng khác cụm trong kết quả dự đoán.

Chỉ số FMI dao động từ 0 đến 1, với 1 biểu thị sự trùng khớp hoàn hảo và 0 biểu thị không có sự tương đồng ngoài ngẫu nhiên.

Chỉ số Rand Index (RI):

Rand Index đo lường sự tương đồng giữa hai phân cụm bằng cách so sánh tất cả các cặp điểm dữ liệu và kiểm tra xem chúng có được gán vào cùng một cụm hoặc khác cụm trong cả hai phân cụm hay không. Công thức RI như sau:

$$RI = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{2.25}$$

Tuy nhiên, RI không tính đến sự trùng khớp ngẫu nhiên, điều này có thể gây sai lệch khi số cụm nhiều hoặc dữ liệu mất cân bằng. Do đó, Chỉ số Rand Điều chỉnh (Adjusted Rand Index - ARI) được sử dụng để hiệu chỉnh theo sự trùng khớp kỳ vọng.

$$ARI = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{\max(\mathrm{RI}) - \mathbb{E}[\mathrm{RI}]} \tag{2.26}$$

Trong đó:

- $\mathbb{E}[\mathrm{RI}]$ là giá trị RI kỳ vọng theo giả thuyết độc lập.
- $\max(\mathrm{RI})$ là giá trị RI cực đại có thể đạt được.

Chỉ số ARI có giá trị từ -1 đến 1, trong đó 1 biểu thị sự trùng khớp hoàn hảo, 0 biểu thị trùng khớp ngẫu nhiên, và giá trị âm biểu thị phân cụm còn tệ hơn ngẫu nhiên.

Bảng 2.1 so sánh các chỉ số đánh giá của bài toán phân cụm.

Bảng 2.1: So sánh các chỉ số đánh giá phân cụm

Chỉ số	Mục đích	Thang đo	Ưu điểm	Hạn chế
MI	Đo lường mức độ phụ thuộc giữa kết quả phân cụm và nhãn thực	$[0,+\infty)$	Trực quan, phản ánh thông tin chia sẻ	Dễ bị thiên lệch khi số cụm nhiều
AMI	Hiệu chỉnh MI theo xác suất ngẫu nhiên	$[-1,1]$	Ổn định, không phụ thuộc số cụm	Tính toán phức tạp hơn MI
NMI	Chuẩn hóa MI theo entropy	$[0,1]$	So sánh được giữa nhiều bộ dữ liệu khác nhau	Có thể đánh giá cao các phân cụm không cân bằng
CS	Kiểm tra các điểm cùng nhãn thực có được gom đúng cụm	$[0,1]$	Đơn giản, dễ hiểu	Không phản ánh hết độ phân tách
FMI	Đo độ tương đồng theo cặp điểm (precision–recall)	$[0,1]$	Cân bằng giữa precision và recall	Nhạy cảm với dữ liệu mất cân bằng
RI	So sánh cặp điểm có cùng/khác cụm	$[0,1]$	Dễ tính toán	Không hiệu chỉnh theo ngẫu nhiên
ARI	Hiệu chỉnh RI theo ngẫu nhiên	$[-1,1]$	Đánh giá công bằng hơn RI	Có thể khó diễn giải trực quan

2.3 Học đặc trưng không giám sát

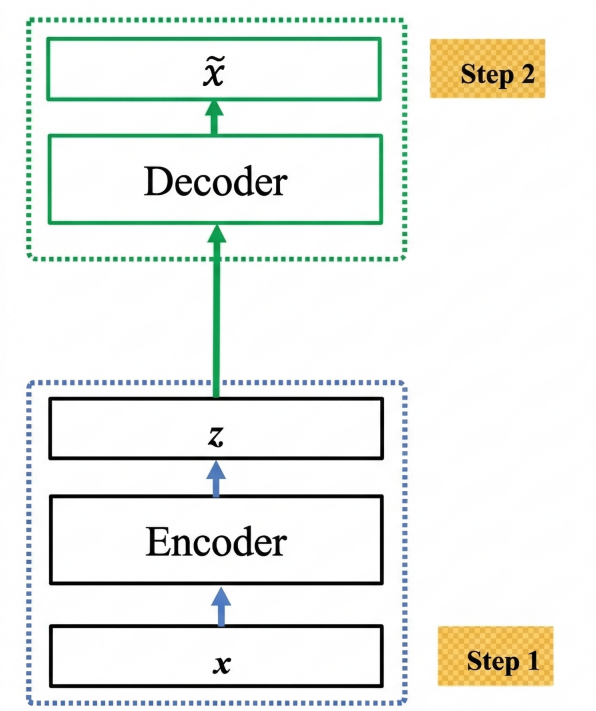
Trong điều kiện dữ liệu thường khan hiếm nhãn, các kỹ thuật học sâu không giám sát đóng vai trò then chốt để trích xuất các đặc trưng tiềm ẩn. Phần này trình bày cơ sở của hai mô hình nền tảng: Mạng nơ-ron tự mã hóa (AE) và tự mã hóa biến phân (VAE).

2.3.1 Mạng nơ-ron tự mã hóa (Autoencoder - AE)

AE là mô hình học cách tái tạo dữ liệu đầu vào qua một biểu diễn nén, gồm hai thành phần:

- **Bộ mã hóa (Encoder):** Ánh xạ đầu vào x sang không gian tiềm ẩn $y = s_f(Wx+b)$.
- **Bộ giải mã (Decoder):** Tái tạo tín hiệu gốc từ y thành $z = s_g(W'y+b')$.

Trong đó, W, W' là ma trận trọng số, b, b' là hệ số tự do, và s_f, s_g là các hàm kích hoạt. Mục tiêu huấn



Hình 2.10: Sơ đồ nguyên lý hoạt động của mạng AE.

luyện là tìm tập tham số θ để cực tiểu hóa sai số tái tạo:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum L(x,z) \tag{2.27}$$

Tùy vào tính chất dữ liệu [41], hàm mất mát L có thể là sai số bình phương trung bình (MSE) đối với dữ liệu thực liên tục:

$$L_{MSE}(x,z) = \frac{1}{2}\sum_i (x_i - z_i)^2 \tag{2.28}$$

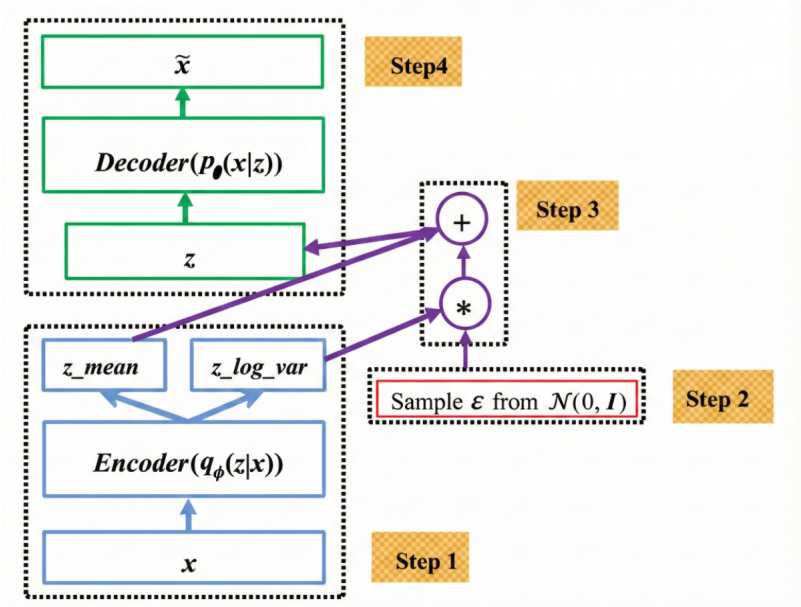
Hoặc hàm Entropy chéo (Cross-Entropy) nếu dữ liệu thuộc khoảng $[0,1]$ hoặc có phân phối Bernoulli:

$$L_{CE}(x,z) = -\sum_i [x_i \log z_i + (1-x_i)\log(1-z_i)] \tag{2.29}$$

2.3.2 Mạng nơ-ron tự mã hóa biến phân (VAE)

VAE [36] cải tiến AE bằng cách ánh xạ đầu vào thành một phân phối xác suất (thường là phân phối Gauss đa chiều), giúp không gian tiềm ẩn liên tục và mang ý nghĩa ngữ nghĩa cao hơn (Hình 2.11).

Để mạng có thể lan truyền ngược, một vector đặc trưng z được lấy mẫu thông qua kỹ thuật tái tham số



Hình 2.11: Sơ đồ nguyên lý hoạt động của mạng VAE.

hóa (reparameterization trick):

$$z = \mu + \varepsilon \cdot \exp\left(\frac{1}{2}\log\sigma^2\right) \tag{2.30}$$

Với μ và $\log\sigma^2$ là trung bình và log phương sai do bộ mã hóa dự đoán, $\varepsilon \sim \mathcal{N}(0,I)$ là nhiễu chuẩn tắc.

Để tăng cường khả năng phân tách dữ liệu trong không gian đặc trưng, các nghiên cứu [42, 43] đề xuất kết hợp hàm mất mát tái tạo với hàm tối ưu khoảng cách (như Triplet Loss):

$$L_{total} = L_{rec} + L_{KL} + L_{triplet} \tag{2.31}$$

Trong đó L_{rec} là sai số tái tạo, L_{KL} là phân kỳ Kullback-Leibler ép phân phối tiềm ẩn gần với phân phối chuẩn, và $L_{triplet}$ giúp gom nhóm các mẫu cùng lớp và đẩy xa các mẫu khác lớp.

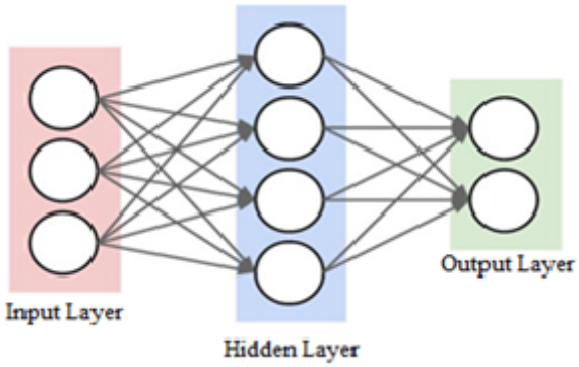
2.4 Cơ sở lý thuyết về phân loại tín hiệu

2.4.1 Mạng Nơ-ron nhân tạo cho bài toán phân loại

Mạng neuron (NN) là một công cụ mạnh mẽ cho các nhiệm vụ phân loại. Như thể hiện trong Hình 2.12, NN là sự kết hợp của nhiều lớp perceptron. Thông thường, có ba loại lớp có thể được tìm thấy trong NN. Lớp đầu vào là lớp ngoài cùng bên trái của mạng và nó đại diện cho các đầu vào

của mạng. Số lượng các nút trong lớp đầu vào là số lượng các đặc trưng. Lớp đầu ra là lớp dưới cùng bên phải của mạng đại diện cho các đầu ra của mạng. Số lượng nút trong lớp đầu ra là số lớp mà chúng ta muốn phân loại. Một lớp ẩn là một lớp giữa đầu vào và lớp ra đại diện cho suy luận logic của mạng. Một NN có thể có nhiều lớp ẩn; do đó mô hình đại diện của một NN có thể được tìm thấy trong phương trình (2.32)

$$y(x,W) = f_L(f_{L-1}(f_{L-2}(,,,(f_2(f_1(x))))))) \tag{2.32}$$



Hình 2.12: Sơ đồ khối một mạng Neuron

Ở đây, x_i là mẫu thứ i^{th} của tập dữ liệu, w là trọng số của mạng và $f_l(x)$ là đặc trưng được trích ở lớp l^{th}. Ký hiệu σ_l là hàm kích hoạt và W_l là trọng số của lớp thứ l^{th} tương ứng; lớp tương ứng $f_l(,)$ của mạng NN được trình bày trong công thức (2.33)

$$f_l(x) = \sigma_l(W_l f_{l-1}(x)) \tag{2.33}$$

Để huấn luyện mạng NN, chúng tôi mong muốn rằng công việc dự đoán mẫu $y(x_i,W)$ thứ i^{th} phải tương tự với mục tiêu t_i. Trong một nhiệm vụ phân loại, sự kiện tương tự có thể được mô hình hóa bằng hàm mất mát cross-entropy trong công thức (2.34). Hàm mất mát nhỏ hơn nghĩa là giống nhau cao hơn.

$$J_{class} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} -t_{i,j} log(P(y(x_i,W) = j)) \tag{2.34}$$

Với t_i là vectơ nhãn của mẫu thứ i^{th} và $P(y(x_i,W) = j)$ là xác suất dự đoán $y(x_i,W)$ thuộc mẫu j^{th}.

Thông thường, xác suất này là đầu ra của mạng NN.

Khi số lượng nút ẩn được tăng lên, NN sẽ dày và phức tạp hơn. Để tránh bị overfitting trong một mạng dày, hàm mất mát chuẩn hóa đã được đưa ra để bỏ qua các đặc tính ít quan trọng hơn. Ký hiệu L là số lớp và s_l là số nút trong lớp thứ l^{th}, hàm mất mát chuẩn hóa chính quy được trình bày trong công thức (2.35).

$$J_{Regu} = \frac{1}{N} \sum_{l=1}^{L-1} \sum_{i=1}^{S_l} \sum_{i=1}^{S_{l+1}} W_l^{i,j} \tag{2.35}$$

Hàm mất mát J_{Regu} là nhỏ nhất khi tất cả $W_l^{i,j}$ bằng không. Trong mẫu này, mạng không thể học được bất cứ điều gì. Do đó, một tham số λ được sử dụng để kiểm soát sự đóng góp của giới hạn chính quy hóa vào quá trình dự đoán dưới dạng công thức (2.36).

$$J_{train} = J_{class} + \lambda J_{Regu} \tag{2.36}$$

Để tối ưu hóa hàm mất mát J_{train} có thể dùng thuật toán Gradient Descent để cập nhập trong số:

$$W^{\tau+1} = W^{\tau} - \alpha \frac{\Delta J_{train}}{\Delta W} \tag{2.37}$$

2.4.2 Các chỉ số đánh giá cho bài toán phân loại đối tượng

Khi xây dựng xong một mô hình phân loại chúng ta cần đánh giá hiệu quả sử dụng của mô hình và so sánh chúng với các mô hình khác. Các phương pháp thường được sử dụng là: độ chính xác, ma trận nhầm lẫn, đường cong ROC, khả năng dự đoán mẫu dương tính (precision), độ nhạy (recall), và F1 score.

Ma trận nhầm lẫn

Ma trận nhầm lẫn là phương pháp có thể giúp chỉ ra cụ thể từng loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất và dữ liệu thuộc lớp nào thường bị phân loại nhầm lẫn vào lớp khác. Confusion Matrix có thể có nhiều điểm dữ liệu thực sự thuộc về một lớp và được dự đoán là rơi vào một lớp. Confusion matrix có dạng sau:

Ma trận nhầm lẫn được xây dựng dựa trên bốn khái niệm chính:

1. True Positive (TP): Số lượng mẫu dương được mô hình phân loại đúng.

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Hình 2.13: Ma trận nhầm lẫn.

2. False Positive (FP): Số lượng mẫu âm bị mô hình phân loại sai (mẫu âm bị dự đoán là dương).
3. True Negative (TN): Số lượng mẫu âm được mô hình phân loại đúng.
4. False Negative (FN): Số lượng mẫu dương bị mô hình phân loại sai (mẫu dương bị dự đoán là âm)

Bằng cách phân loại mẫu dữ liệu vào các phần tử của confusion matrix, có thể tính toán các chỉ số khác nhau như precision, recall, accuracy và F1 score để đánh giá hiệu suất của mô hình phân loại. Confusion matrix cung cấp một cái nhìn tổng quan về khả năng phân loại của mô hình và giúp xác định các lỗi phân loại cụ thể mà mô hình đang mắc phải.

Độ chính xác

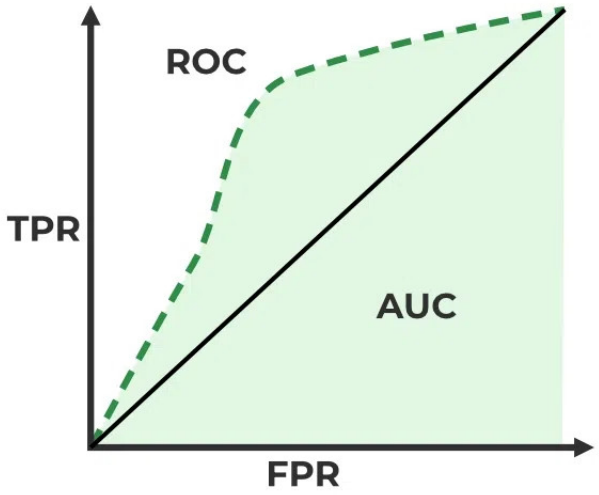
Accuracy score tính tỷ lệ giữa số điểm được dự đoán đúng trên tổng số điểm trong dữ liệu thử nghiệm. Phương pháp này đơn giản nhưng chỉ biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng mà không chỉ ra cụ thể cho mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhất và dữ liệu thuộc lớp nào nào thường bị nhầm lẫn vào các lớp khác nhau. Accuracy được tính theo công thức sau:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.38}$$

Đường cong ROC

Đường cong Receiver Operating Characteristic (ROC) là một biểu đồ đường cong được sử dụng để đánh giá hiệu suất của mô hình phân loại trong các bài toán nhị phân. Trên ROC curve, trục hoành thể hiện tỷ lệ False Positive Rate (FPR), còn trục tung thể hiện tỷ lệ True Positive Rate (TPR), còn được gọi là recall. Mỗi điểm trên ROC curve tương ứng với một ngưỡng (threshold) khác nhau trong quá trình quyết định phân loại. mô hình phân loại tạo ra các dự đoán xác suất và dựa trên ngưỡng để

quyết định nhãn cuối cùng cho mỗi mẫu. Bằng cách thay đổi ngưỡng này, ta có thể tính toán FPR và TPR tương ứng, từ đó vẽ được đường cong ROC.



Hình 2.14: Chỉ số đánh giá phân loại ROC-AUC

Đường cong ROC cung cấp một cái nhìn tổng quan về khả năng phân loại của mô hình phân loại và đo lường khả năng phân biệt giữa các lớp. Đường cong càng nằm gần góc trên bên trái, mô hình càng có khả năng phân loại tốt hơn.

Diện tích dưới ROC curve (AUC-ROC) là một thước đo số liệu quan trọng của hiệu suất phân loại. Nó đại diện cho khả năng mô hình phân loại xếp hạng đúng giữa các mẫu dương và âm. Nếu AUC-ROC gần 1, mô hình có hiệu suất tốt hơn trong việc phân loại.

Diện tích dưới đường cong AUC

Area Under the Curve (AUC) đề cập đến diện tích dưới đường cong của đồ thị tỷ lệ đúng dương (True Positive Rate, TPR) theo tỷ lệ đúng âm (False Positive Rate, FPR) khi áp dụng một mô hình phân loại.

Trong một bài toán phân loại, điểm mô hình được sắp xếp theo thứ tự từ cao đến thấp. AUC đo lường khả năng của mô hình phân loại phân biệt giữa các điểm dương và âm. Nó cho biết xác suất một mẫu dương được phân loại đúng cao hơn so với xác suất một mẫu âm được phân loại đúng. AUC có giá trị từ 0 đến 1, trong đó 0 đại diện cho mô hình phân loại tệ nhất (dự đoán sai hoàn toàn) và 1 đại diện cho mô hình phân loại hoàn hảo (dự đoán đúng hoàn toàn). AUC thường được sử dụng để so sánh hiệu suất giữa các mô hình phân loại khác nhau và đánh giá khả năng phân loại của một mô hình. Nếu AUC càng gần 1, thì mô hình càng có khả năng phân loại tốt hơn.

Chương 3

PHÁT HIỆN TÀU BIỂN TỪ DỮ LIỆU ẢNH

Chương 3 tập trung nghiên cứu bài toán phát hiện tàu biển dựa trên dữ liệu ảnh camera. Các mô hình học sâu hiện đại được phân tích và lựa chọn làm nền tảng xây dựng hệ thống. Trên cơ sở đó, luận án đề xuất phương pháp cải tiến thông qua cơ chế lựa chọn đặc trưng nhằm nâng cao khả năng khái quát, đặc biệt trong điều kiện dữ liệu huấn luyện hạn chế. Bên cạnh kiến trúc dựa trên CNN, mô hình phát hiện đối tượng dựa trên Transformer cũng được triển khai và đánh giá. Các thí nghiệm được thực hiện trên nhiều cấu hình khác nhau để phân tích ảnh hưởng của siêu tham số, vị trí chèn mô-đun và kích thước dữ liệu. Kết quả cho thấy phương pháp đề xuất đạt hiệu năng cao và ổn định hơn so với các phương pháp tham chiếu.

3.1 Phát hiện tàu biển bằng mô hình Transformer

Các bộ phát hiện đối tượng hiện tại thường giải quyết bài toán phát hiện đối tượng theo các cách gián tiếp như các phương pháp hai giai đoạn tạo ra các vùng đề xuất sau đó tiến hành phát hiện và phân loại đối tượng trên các vùng này, hay các mô hình YOLO đưa ra dự đoán về hộp bao và phân loại đối tượng dựa trên các điểm neo được thiết lập trước. Hiệu suất của những mô hình này bị ảnh hưởng đáng kể bởi các bước hậu xử lý để tinh chỉnh các kết quả đạt được, loại bỏ các phát hiện trùng lặp hay việc thiết kế các điểm neo. Đối với mô hình YOLO ta còn cần thiết lập một số tham số chẳng hạn như tọa độ các điểm neo nhưng DETR có thể đưa ra dự đoán tập hợp một cách trực tiếp (end-to-end) mà không cần phải thiết lập các tham số ban đầu như điểm neo cũng như không cần sử dụng đến kỹ thuật hậu xử lý non-max suppression. DETR hay Detection Transformer với một kỹ thuật được đơn giản hóa, đưa ra các dự đoán thông qua kỹ thuật làm khớp hai hướng (Bipartite Matching) và một kiến trúc mã hóa- giải mã (Encoder - Decoder) theo cơ chế của mô hình Transformer [44].

3.2 Các phương pháp phát hiện tàu biển

Dựa trên phương pháp phát hiện đối tượng truyền thống, một số tuỳ biến đã được giới thiệu nhằm cải thiện hiệu suất của bộ phát hiện tàu. Liu_2022 [45] dựa trên khung sườn của mô hình SSD [15] và bộ rút trích đặc trưng VGG để phát hiện một con tàu ở quy mô nhỏ. Tác giả [45] đã sử dụng cơ chế chú ý (attention) cục bộ để hợp nhất chéo các đặc trưng; đồng thời, một mô-đun kết hợp các đặc trưng từ các thang đo khác nhau nhằm cải thiện kết quả phát hiện. Các phiên bản khác nhau của YOLO cũng được nhiều công trình sử dụng để tăng cường khả năng phát hiện trên bộ dữ liệu tàu. Dựa trên mô hình YOLO, Biaohua_2022 [46] đã giới thiệu "Mạng nhất quán tỷ lệ và chú ý liên cấp" (CARC) để phát hiện tàu. Trong bài báo này, bộ rút trích đặc trưng là Resnet-34; Khối kết hợp (neck) là một mô-đun nhiều cấp độ sử dụng kỹ thuật attention theo kênh và kỹ thuật attention không gian để trích xuất các đặc trưng ở các tỷ lệ khác nhau. Các đặc trưng được nối và đưa vào các đầu phân loại truyền thống. Cui_2019 [47], Liu_2020 [48] và Li_2021 [49] dựa trên YOLOV3 để phát hiện tàu. Cui_2019 [47] đã giới thiệu YOLOv3-ship, bao gồm các kỹ thuật cụm để lựa chọn các mỏ meo. Đồng thời với đó là các cải tiến mô hình khi sử dung kỹ lựa chọn đặc trưng khong kiến trúc của mô hình phát hiện đối tượng. Liu_2020 [48] giới thiệu hai phương pháp cài đặt mỏ neo mới và kết hợp khối rút trích dữ liệu đa độ phân giải để nâng cao hiệu suất của YOLOv3. Thay vì sử dụng FPN [50] để kết nối bộ rút trích đặc trưng với bộ phân loại, phương pháp này đã sử dụng phương pháp Cross PANet, có thể kết hợp các thông tin cấp thấp về vị trí của đối tượng với thông tin cấp cao về ngữ nghĩa của đối tượng. Li_2021 [49] dựa trên YOLOv3 Tiny [51] để phát triển một quy trình huấn luyện hai bước. Ở đây, khối CBAM [52] được sử dụng để phát hiện các mục tiêu lớn; sau đó, việc tinh chỉnh được thực hiện trên mô hình để phát hiện các mục tiêu nhỏ.

Gần đây, các phiên bản cải tiến của mô hình YOLO đã được giới thiệu để phát hiện tàu. Zhang_2021 [53] đã sử dụng YOLOv4 với Tích chập phân tách theo chiều sâu ngược (RDSC) để phát hiện tàu. RDSC được đề xuất đã thay thế Tích chập phân tách theo chiều sâu (DSC) [54] trong ResUnit của mạng YOLOv4. Với sự trợ giúp của RDSC, độ phức tạp của mô hình mạng được giảm bớt mà vẫn đảm bảo độ chính xác. Han_2021 [55] cũng sử dụng bộ rút trích đặc trưng của YOLOV4 với cơ chế chú ý để cải thiện hiệu suất. Light_SDNet [56] đã sửa đổi mô hình YOLO5 bằng khối Gost [57] và DepthWise Convolution (DWConv) [58] để giảm số lượng tham số; Ngoài ra, việc tăng cường dữ liệu như tạo sương mù và tạo mưa đã được giới thiệu để làm phong phú thêm tập huấn luyện. Gần đây, YOLOX đã được coi là phương pháp mạnh mẽ và hiệu quả để phát hiện đối tượng; Zhang_2022 [59]

đã dựa trên mô hình YOLOX để thiết kế một phương thức có trọng số nhẹ. Thay vì sử dụng PANnet [60] để hợp nhất các đặc trưng, bài báo đã sử dụng mộ bộ kết hợp các tham số (LACFF) để khắc phục hiện tượng không nhất quán về mặt tỷ lệ giữa các bản đồ đặc trưng. Các đặc trưng của tất cả các lớp khác được điều chỉnh theo cùng một hình dạng. Sau đó, các kênh được hợp nhất theo các trọng số đã học. Tương tự như Zhang_2022 [59], công trình của Nghiên cứu sinh cũng dựa trên YOLOX; tuy nhiên, nghiên cứu này không tập trung vào việc phối hợp các đặc trưng mà đưa ra một hàm mất mát cho phép chọn các đặc trưng phù hợp trên nhánh phân loại.

Các phương pháp dựa trên mô hình transformer[18] cũng là một giải pháp khả thi để phát hiện tàu. Yani_2022 [61] đã sử dụng phương pháp học chắt lọc để huấn luyện bộ phát hiện tàu dựa trên nền tảng của phương pháp DETR. Một mô hình giáo viên được huấn luyện dựa trên Bộ dữ liệu CoCo quy mô lớn, và mô hình học học theo được tinh chỉnh dựa trên bộ dữ liệu Seaship [62]. Phương pháp này giúp giảm FLOP và số lượng tham số. Tuy nhiên, chỉ số mAP của nó không được cải thiện so với phương pháp DETR thông thường.

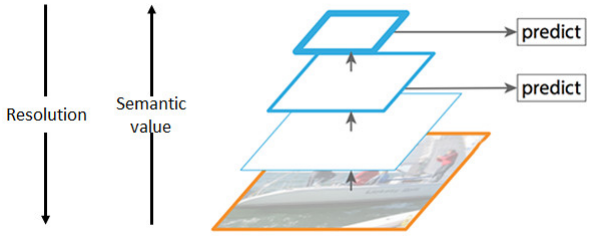
3.3 Phương pháp phát hiện tàu dựa trên mạng YOLOX kết hợp VIB

Như đã trình bày chi tiết về kiến trúc nền tảng của họ mô hình YOLO, đặc biệt là biến thể YOLOX tại Chương 2, mô hình này cho thấy ưu điểm vượt trội về tốc độ và khả năng nhận dạng. Tuy nhiên, đối với bài toán cảnh giới bờ biển, đặc thù các đối tượng tàu thường ở xa, kích thước nhỏ và chịu nhiều nhiễu nền từ mặt biển, cấu trúc gốc của YOLOX vẫn còn hạn chế trong việc trích xuất và tinh giản đặc trưng. Để giải quyết vấn đề này, thay vì trình bày lại kiến trúc gốc, phần này sẽ đi sâu vào cấu trúc cải tiến do luận án đề xuất: tích hợp mô-đun VIB vào mạng YOLOX để nâng cao khả năng chọn lọc đặc trưng đối tượng.

FPN (Feature Pyramid Networks)

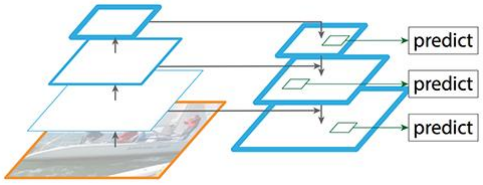
Một trong những thách thức kinh điển của bài toán phát hiện đối tượng đó là các đối tượng có thể xuất hiện với nhiều kích thước khác nhau. Đặc biệt việc phát hiện vật thể nhỏ được coi là thách thức lớn. Việc phát hiện các đối tượng có kích thước nhỏ là một vấn đề đáng được giải quyết để nâng cao độ chính xác. FPN là mô hình mạng được thiết kế ra dựa trên trên khái niệm kim tự tháp để giải quyết vấn đề này.

FPN bao gồm một đường từ dưới lên (bottom-up pathway) và một đường từ trên xuống (top-down pathway), trong khi các thuật toán khác chỉ sử dụng đường từ dưới lên (bottom-up). Đường từ dưới lên là một mạng CNN thông thường dùng để trích xuất các đặc trưng được minh họa như Hình 3.1. Càng lên cao, độ phân giải càng giảm (kích thước của bản đồ đặc trưng giảm), và giá trị thông tin về



Hình 3.1: Hình minh họa Đường dẫn từ dưới lên

ngữ cảnh càng cao. mô hình đưa ra quyết định dựa vào nhiều bản đồ đặc trưng. Nhưng tầng (layer) ở dưới không được sử dụng để phát hiện đối tượng vì những này có độ phân giải cao nhưng giá trị ngữ nghĩa của chúng lại không đủ cao. Do đó nó chỉ sử dụng những lớp ở bên trên để phát hiện đối tượng dẫn đến mô hình hoạt động không tốt khi phát hiện các đối tượng nhỏ.

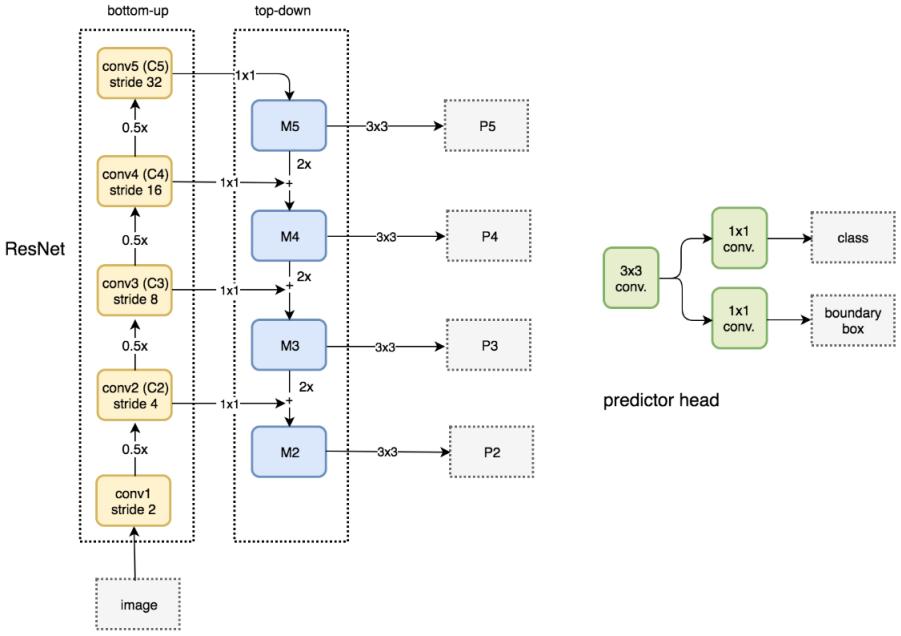


Hình 3.2: Cấu trúc lan truyền từ dưới lên và từ trên xuống của FPN

FPN xây dựng thêm mô hình từ trên xuống, nhằm mục đích xây dựng các lớp có độ phân giải cao từ các tầng có ngữ nghĩa cao được minh họa ở Hình 3.2. Trong quá trình xây dựng lại các lớp từ cao xuống thấp, chúng ta sẽ gặp một thách thức là bị mất mát thông tin của các đối tượng. Ví dụ một đối tượng nhỏ khi lên các lớp ở trên cao sẽ bị biến mất; và từ các lớp trên cao khi đi ngược lại các lớp bên dưới sẽ không thể tái tạo lại đối tượng nhỏ đó. Để giải quyết vấn đề này, người ta sử dụng tạo các kết nối giữa các lớp liền kề khi đi ngược từ lớp cao xuống các lớp thấp. Các bản đồ đặc trưng ở các lớp cao giúp tăng cường thông tin ngữ nghĩa, và các đặc trưng từ các lớp thấp vẫn còn lưu lại thông tin của các đối tượng nhỏ. Việc kết hợp thông tin ngữ nghĩa từ các lớp cao và thông tin chi tiết từ các lớp thấp giúp bộ phát hiện đối tượng dự đoán các vị trí của đối tượng thực hiện tốt hơn (hạn chế tốt nhất việc mất mát thông tin).

Hình 3.3 bên dưới mô tả chi tiết đường đi theo từ dưới lên và từ trên xuống. P2, P3, P4, P5 là các kim

tự tháp của các bản đồ đặc trưng. Đường đi từ dưới lên sử dụng mô hình ResNet. Nó chứa rất nhiều



Hình 3.3: Minh họa đường đi theo từ dưới lên và từ trên xuống

khối nhân chập (Convi với i = 1...5), mỗi khối có một vài lớp nhân chập. Theo chiều mũi tên đi lên, độ phân giải theo miền không gian giảm một nửa (do tham số stride tăng lên gấp đôi). Đầu ra của mỗi khối nhân chập được đánh số là C_i và sẽ được sử dụng lại trong chiều đi từ trên xuống (top-down). Các nhà nghiên cứu dùng dùng phép nhân chập kích thước $1x1$ để giảm số kênh đặc trưng của C_5 xuống 256-d để tạo ra đặc trưng M_5. M_5 là bản đồ đặc trưng đầu tiên được sử dụng cho nhánh dự đoán P_5 để tạo ra kết quả phát hiện đối tượng. Theo chiều mũi tên đi xuống của nhánh từ trên xuống, chúng ta sử dụng kỹ thuật phóng to ảnh (upsampling) để tăng kích thước của M_5 lên 2 lần bằng cách sử dụng kỹ thuật lấy mẫu dựa trên những điểm gần nhất. Sau đó lại áp dụng nhân chập với tỷ lệ 1x1 cho bản đồ đặc trưng C_4 rồi cộng chúng lại với nhau để nhận được bản đồ đặc trưng M_4. Sau đó áp dụng lớp nhân chập 3x3 cho M_4 để nhận được P_4. Chúng ta lặp lại quá trình này cho P_3, P_2. Chúng ta chỉ dừng lại ở P_2 mà không tới P_1 do độ phân giải của C_1 quá lớn, điều này sẽ làm giảm tốc độ xử lý. Bởi vì chúng ta chia sẻ cùng nhánh phân loại và nhánh hồi quy phát hiện vị trí cho mỗi ngõ ra của bản đồ đặc trưng, do đó tất cả bản đồ đặc trưng dạng kim tự tháp (P5, P4, P3, P2) đều có phải chuẩn hóa về 256 kênh.

Ở đây, mô hình FPN không phải là mô hình phát hiện đối tượng hoàn chỉnh. Nó chỉ là mô hình trích xuất đặc trưng và được sử dụng cùng với một thuật toán phát hiện đối tượng. Các bản đồ đặc trưng từ P_2 đến P_5 trong hình bên dưới độc lập với nhau. Một phép nhân chập 3×3 trượt lần lượt qua các bản đồ này và trích xuất các đặc trưng cần thiết. Sau đó các đặc trưng được đưa qua một phép nhân chập

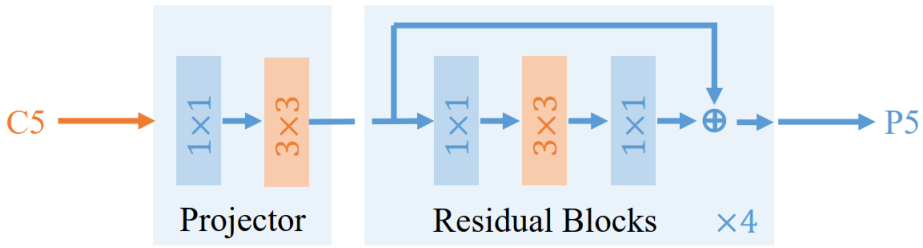
1×1 để tính kết quả quá hiện vị trí đối tượng và một phép nhân chập 1×1 để tính kết quả phân loại.

Ban đầu, FPN được tạo ra để thay thế các bộ trích đặc trưng được sử dụng như trong Faster R-CNN và tạo ra bản đồ đặc trưng nhiều kích thước với chất lượng thông tin tốt hơn bộ lọc thông thường. Tuy nhiên nó đã được mở rộng để có thể làm việc ở nhiều kiến trúc khác nhau.

Bộ mã hoá dựa trên phép nhân chập giãn cách (Dilated Encoder)

Nhận biết các đối tượng ở các tỉ lệ rất khác nhau là một thách thức cơ bản trong việc phát hiện đối tượng. Một giải pháp khả thi cho thách thức này là tận dụng các đặc trưng nhiều cấp (multiple-level features). Trong các bộ mã hóa nhiều ngõ vào nhiều ngõ ra (cơ chế FPN), các đặc trưng nhiều cấp được xây dựng để trích thông tin từ các các trường tiếp nhận (receptive fields) khác nhau, và phát hiện các đối tượng ở cấp có trường tiếp nhận phù hợp với thang đo của chúng. Tuy nhiên đối với bộ giải mã một ngõ vào và một ngõ ra, đây chính là thách thức lớn bởi vì nó chỉ có một đặc trưng đầu ra với trường tiếp nhận là một hằng số. Trường tiếp nhận của đặc trưng chỉ có thể bao phủ một phạm vi tỷ lệ giới hạn, dẫn đến hiệu suất kém nếu tỷ lệ của đối tượng không khớp với trường tiếp nhận. Để đạt được mục tiêu phát hiện tất cả các đối tượng bằng một bộ giải mã một ngõ vào một ngõ ra (SISO encoder), chúng ta phải tìm cách tạo ra một đặc trưng duy nhất những có các trường tiếp nhận khác nhau, bù đắp cho việc thiếu các đặc trưng ở các độ phân giải khác nhau [1].

Kỹ thuật mã hóa dựa trên phép nhân chập giãn cách (Dilated Encoder) được thiết kế dựa trên mục tiêu xây dựng một bộ rút trích đặc trưng chỉ có một ngõ vào và một ngõ ra nhưng có thể trích xuất thông tin ở nhiều mức độ nhận thức khác nhau. Kiến trúc của bộ giải mã này được mô tả như trong Hình 3.4. Dilated Encoder gồm hai thành phần chính là khối "projector" và các khối "Residual Block". Trước



Hình 3.4: Cấu trúc của Dilated Encoder [1]

tiên, lớp Projector áp dụng một lớp tích chập 1 × 1 để giảm kích thước kênh, sau đó thêm một lớp tích chập 3 × 3 để tinh chỉnh ngữ cảnh, giống như trong FPN. Sau đó xếp chồng bốn khối "Dilated residual blocks" liên tiếp với các tỷ lệ giãn nở khác nhau trong các lớp tích chập 3 × 3 để tạo ra các

đặc trưng đầu ra với nhiều trường tiếp nhận khác nhau, bao gồm tất cả các tỉ lệ của đối tượng. Tất cả các lớp tích chập trong "Residual Blocks" được theo sau bởi một lớp chuẩn hóa "batchnorm" và một hàm kích hoạt "ReLU". Trong khi đó, khối projector chỉ sử dụng các lớp tích chập thông thường và lớp chuẩn hóa dữ liệu "batchnorm".

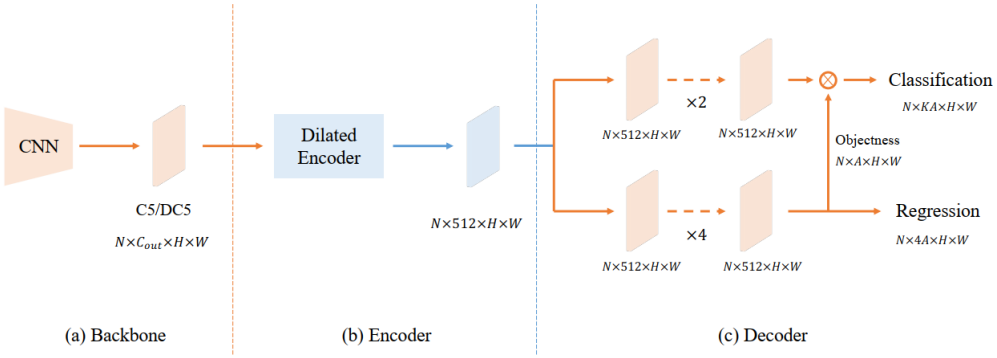
Kỹ thuật loại bỏ đối tượng trùng lắp khi áp dụng tại nhiều nhánh khác nhau(Uniform Matching)

Trong các mô hình phát hiện đối tượng dựa trên hộp neo, các chiến lược để xác định các đường bao dương tính bị chi phối bởi việc đo IoU giữa các điểm neo và đường bao được gắn nhãn. Trong RetinaNet, nếu IoU tối đa của điểm neo và đường bao được gắn nhãn lớn hơn ngưỡng 0,5, thì hộp neo này sẽ được đặt là dương tính thật (tức là phát hiện đúng). Ta gọi nó là Max-IoU matching. Trong bộ giải mã nhiều ngõ vào nhiều ngõ ra (MiMo encoder), các hộp neo được xác định trước trên nhiều cấp và các đường bao chuẩn tạo ra các hộp neo dương tính ở các cấp đặc trưng tương ứng với tỷ lệ của chúng. Với cơ chế chia nhỏ để tính toán, kỹ thuật làm khớp Max-IoU cho phép các đường bao chuẩn trong mỗi tỷ lệ tạo ra đủ số lượng hộp neo dương tính (hộp neo gắn với một đối tượng). Tuy nhiên, khi sử dụng bộ giải mã vào đơn ra đơn (SiSo encoder), số lượng hộp neo giảm nhiều so với hộp neo trong MiMo encoder, từ 100k xuống còn 5k, dẫn đến hộp neo thưa thớt (có rất ít hộp neo dương tính-các hộp neo được phát hiện trùng với đối tượng). Việc này gây ra thách thức khi gán các đường bao được phát hiện với các đối tượng được gắn nhãn. Các đối tượng có kích thước lớn sẽ tạo ra nhiều hộp neo dương tính hơn các đối tượng có kích thước nhỏ. Điều này gây ra vấn đề mất cân bằng cho các hộp neo dương tính. Sự mất cân bằng này làm cho các bộ phát hiện chú ý đến các đối tượng có khích thước lớn trong khi bỏ qua các đối tượng có kích thước nhỏ trong khi huấn luyện.

Chiến lược Uniform Matching đã được đề xuất để giải quyết vấn đề mất cân bằng này trong các hộp neo dương tính. Uniform Matching sử dụng k hộp neo gần nhất để trở thành các hộp neo dương tính cho mỗi đối tượng được gán nhãn. Điều này đảm bảo rằng tất cả các hộp neo được gán nhãn có thể được khớp với cùng số lượng hộp neo dương tính một cách đồng nhất bất kể kích thước của chúng. Sự cân bằng trong các mẫu dương tính đảm bảo rằng tất cả các đường bao được gắn nhãn đều tham gia vào quá trình huấn luyện và đóng góp như nhau. Bên cạnh đó theo ta có thể đặt ngưỡng để loại bỏ các hộp neo âm tính có IoU lớn (>0,7) và các dương tính hộp neo có IoU nhỏ (<0,15) [1].

Cấu trúc của YOLOF

Cấu trúc của YOLOF gồm 3 phần: mạng cơ sở(backbone), bộ mã hóa(encoder) và bộ giải mã(decoder) được mô tả như Hình 3.5 Backbone: theo bài báo [1] ResNet và ResNeXt được chọn làm của mô hình.



Hình 3.5: Cấu trúc của YOLOF [1]

Đầu ra của mạng cơ sở là bản đồ đặc trưng C5 / DC5 với C_{out} là số lượng kênh của đặc trưng (theo bài báo là 2048 với tỉ lệ downsample là 32). H × W là chiều cao và chiều rộng của bản đồ đặc trưng.

Encoder: sử dụng Dilated Encoder được mô tả ở Hình 3.1. Tương tự như FPN ta thêm hai lớp projector (một tích chập 1 × 1 và một tích chập 3 × 3) sau mạng cơ sở, dẫn đến một bản đồ đặc trưng với 512 kênh. Sau đó, để cho phép đặc trưng đầu ra của bộ mã hóa bao phủ tất cả các đối tượng trên các tỷ lệ khác nhau ta thêm các residual block, bao gồm ba lần chập liên tiếp: phép chập 1 × 1 đầu tiên áp dụng giảm kích thước kênh với tỷ lệ giảm là 4, sau đó là tích chập 3 × 3 với độ giãn nở được sử dụng để phóng to trường tiếp nhận, cuối cùng một phép chập 1 × 1 để khôi phục số kênh.

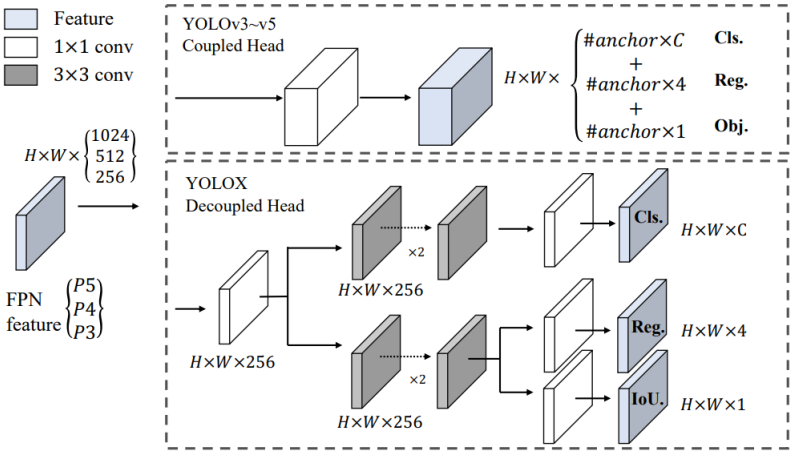
Decoder: ta áp dụng thiết kế chính của RetinaNet, bao gồm hai đầu song song dành riêng cho nhiệm vụ: đầu phân loại (classification head) và đầu hồi quy (regression head) chỉ thêm hai sửa đổi nhỏ. Thứ nhất là đặt số lượng của lớp tích chập ở hai đầu là khác nhau. Ở đầu hồi quy là 4 lớp tích chập theo sau là lớp batch normalization lớp ReLU trong khi ở đầu phân loại chỉ có 2 lớp. Thứ hai là tuân theo Autoassign nhưng thêm dự đoán implicit objectness (không cần giám sát trực tiếp) cho mỗi hộp neo trên đầu hồi quy. Classification score cuối cùng cho tất cả các dự đoán được tạo ra bằng cách nhân kết quả phân loại với implicit objectness tương ứng [1].

YOLOX

Với sự phát triển của phát hiện đối tượng, các mô hình họ YOLO luôn ứng dụng các công nghệ phát hiện tiên tiến nhất có sẵn tại thời điểm đó như sử dụng điểm neo (anchor box) cho YOLOv2 hay

Residual Net cho YOLOv3, để có thể cải thiện tốc độ suy luận và độ chính xác của các ứng dụng thời gian thực. Tuy nhiên, trong hai năm gần đây những tiến bộ lớn trong phát hiện đối tượng đã tập trung vào các bộ phát hiện không có điểm neo, chiến lược gán nhãn nâng cao hay các bộ phát hiện end-to-end (không sử dụng non-max suppression). Những tiến bộ này vẫn chưa được ứng dụng vào các phiên bản YOLOv4, YOLOv5 vì chúng vẫn còn sử dụng điểm neo và các quy tắc gán thủ công để huấn luyện. Và YOLOX đã xuất hiện với những thay đổi bao gồm việc không còn sử dụng hộp neo và thay thế Đầu gộp (đầu gộp) và đầu tách (Decouple head) [2].

Đầu tiên chúng ta sẽ làm rõ sự khác nhau giữa đầu gộp và đầu tách. Mỗi đầu của mạng YOLO sẽ làm hai nhiệm vụ là phân loại và hồi quy. đầu gộp được sử dụng ở mô hình YOLOv3 – YOLOv5 sẽ xử lý hai nhiệm vụ này trên cùng một nhánh. đầu tách được sử dụng trong mô hình YOLOX thì xử lý hai nhiệm vụ này trên hai (hoặc nhiều) nhánh khác nhau cụ thể được mô tả ở Hình 3.6. Một vấn đề với



Hình 3.6: Minh hoạ đầu gộp và đầu tách [2]

đầu gộp là sự xung đột giữa nhiệm vụ phân loại và nhiệm vụ hồi quy. Thách thức này ảnh hưởng đến độ chính xác của mô hình nói chung. Đó là lý do mà hầu hết các bộ phát hiện vật thể, dù là một bước hay hai bước đều sử dụng đầu tách. Để sử dụng đầu tách, ta thực hiện một số thay đổi được minh họa như Hình 3.6. Đối với mỗi cấp của đặc trưng FPN (Feature Pyramid Networks), ta áp dụng lớp conv 1×1 để giảm đố đặc trưng xuống còn 256. Sau đó thêm hai nhánh song song với hai lớp conv 3 ×3 cho mỗi nhiệm vụ phân loại và hồi quy tương ứng. Nhánh IoU đo sự chồng lấn giữa đối tượng và vị trí của lưới được thêm vào nhánh hồi quy [2].

Các thí nghiệm trong bài báo [2] đã chỉ ra rằng việc sử dụng đầu hộp sẽ làm giảm hiệu năng của bộ phát hiện và việc thay thế bằng đầu tách giúp tăng tốc độ hội tụ của mô hình.

Không sử dung hộp neo

Kỹ thuật sử dụng hộp neo được áp dụng từ YOLOv2 đến YOLOv5. Để đạt được hiệu suất phát hiện tối ưu, người ta cần xác định một tập hợp hộp neo tối ưu trước khi huấn luyện bằng cách phân tích phân nhóm dữ liệu. Những hộp neo này phụ thuộc vào miền của dữ liệu (domain-specific) và thiếu tính tổng quát. Thứ hai, cơ chế hộp neo làm tăng độ phức tạp của các đầu phát hiện, cũng như số lượng dự đoán cho mỗi hình ảnh. Việc không sử dụng hộp neo làm giảm số lượng siêu tham số cần phải tinh chỉnh, giảm độ phức tạp của đầu phát hiện đối tượng và loại bỏ những vấn đề liên quan như phân nhóm hộp neo [2].

Để sử dụng kiến trúc YOLO nhưng không bao gồm các hộp neo cũng không quá phức tạp. Ta giảm các số dự đoán cho từng vị trí xuống 1 và để chúng dự đoán trực tiếp bốn giá trị là hai giá trị offsets của góc trên bên trái của lưới, chiều cao và chiều rộng của ô dự đoán. Ta chỉ định vị trí trung tâm của mỗi đối tượng làm mẫu dương tính và xác định trước một phạm vi tỷ lệ để chỉ định mức FPN cho mỗi đối tượng. Việc sửa đổi như vậy làm giảm các thông số và GFLOP của bộ phát hiện đối tượng, làm tăng tốc độ của mô hình nhưng vẫn đạt được hiệu suất tốt hơn [2]. Từ những ưu điểm về khả năng nén đặc trưng của VIB, nghiên cứu này đề xuất tích hợp trực tiếp cơ chế này vào kiến trúc phát hiện đối tượng YOLOX. Cấu trúc chi tiết của mô hình tích hợp này sẽ được trình bày cụ thể trong mục tiếp theo

3.3.1 Lựa chọn đặc trưng bằng VIB

Trước khi trình bày chi tiết về quá trình lựa chọn đặc trưng dựa trên lý thuyết thông tin, luận án thống nhất các thuật ngữ liên quan đến VIB được sử dụng xuyên suốt trong kiến trúc mạng như sau:

- **Mô-đun VIB** (trong một số nghiên cứu còn gọi là *khối VIB*): Là thành phần kiến trúc mạng nơ-ron được thêm vào mô hình để tính toán các tham số thống kê (kỳ vọng μ và phương sai σ^2), qua đó biểu diễn dữ liệu dưới dạng một phân bố xác suất nhằm tạo ra một "nút thắt" thông tin. Để đảm bảo tính nhất quán, luận án sẽ sử dụng duy nhất thuật ngữ **"mô-đun VIB"**.
- **Hàm mục tiêu VIB** (VIB Loss): Là hàm mất mát tổng thể được sử dụng để huấn luyện mô hình, bao gồm hàm mục tiêu của tác vụ chính, và thành phần điều chuẩn dựa trên độ phân kỳ Kullback-Leibler để thực hiện quá trình nén và lựa chọn đặc trưng.

Lựa chọn đặc trưng là quá trình xác định các đặc trưng phù hợp cho một nhiệm vụ cụ thể. Theo lý thuyết thông tin [63], các đặc trưng tốt là những biểu diễn được nén ở mức vừa đủ, đảm bảo giải quyết

được bài toán đặt ra mà không chứa thông tin dư thừa. Yêu cầu này có thể được diễn giải thông qua hai ràng buộc sau:

- Đặc trưng z phải chứa đủ thông tin để dự đoán chính xác biến đầu ra y (trong bài toán này là loại tàu và đường bao đối tượng);
- Từ đặc trưng z, không thể suy ngược lại một cách rõ ràng dữ liệu đầu vào x.

Trong lý thuyết xác suất và lý thuyết thông tin, mức độ phụ thuộc giữa hai biến ngẫu nhiên được đo bằng thông tin liên hợp (mutual information) $I(\cdot)$ [64]. Do đó, hai ràng buộc trên có thể được mô hình hóa bằng cách tối đa hóa $I(y;z)$ và đồng thời giảm thiểu $I(x;z)$. Ràng buộc thứ nhất đảm bảo rằng đặc trưng z mang đủ thông tin để dự đoán nhãn y, trong khi ràng buộc thứ hai đảm bảo rằng z không lưu giữ quá nhiều thông tin về dữ liệu đầu vào x. Với β là hệ số nhân Lagrange, bài toán tối ưu hóa tương ứng được biểu diễn trong phương trình (3.1).

$$L_{IB} = I(y;z) - \beta I(x;z), \tag{3.1}$$

Ở đây $I(y;z)$ và $I(x;z)$ được đại diện bởi phương trình (3.2) và phương trình (3.3), tương ứng.

$$\begin{aligned} I(y;z) &= \int p(y,z) \log \frac{p(y,z)}{p(y)p(z)} dydz = \int p(y,z) \log \frac{p(y \mid z)}{p(y)} dydz \\ &= \int p(y,z) \log p(y \mid z) dydz - \int p(y) \log p(y) dy, \end{aligned} \tag{3.2}$$

$$I(x;z) = \int p(x,z) \log \frac{p(x,z)}{p(x)p(z)} dxdz = \int p(x,z) \log \frac{p(z \mid x)}{p(z)} dxdz, \tag{3.3}$$

Bởi vì các phương trình này không tồn tại các phương pháp giải nhanh để tính toán trực tiếp trong mạng nơ-ron, phương pháp xấp xỉ biến phân được áp dụng. Việc sử dụng các phân bố xấp xỉ này chính là cơ sở toán học hình thành nên phương pháp Nút thắt thông tin biến phân. Cụ thể, giới hạn dưới và giới hạn trên của chúng được sử dụng để tính gần đúng. Phương trình (3.2) và Phương trình (3.3) là các giới hạn trên và giới hạn dưới tương ứng của $I(y;z)$ và $I(x;z)$. Gọi $q(y \mid z)$ là một xấp xỉ cận trên của $p(y \mid z)$, và $q(z)$ là một xấp xỉ cận dưới $p(z)$. Sử dụng đạo hàm phân kỳ KL, giới hạn dưới của phương trình (3.2) được viết lại dưới dạng phương trình (3.4), và giới hạn trên của $I(x;z)$ được viết lại dưới dạng phương trình (3.5).

$$
\begin{aligned}
I(y;z) &\geq \int p(y,z)\log q(y \mid z)dydz - \int p(y)\log p(y)dy \\
&= \int p(z \mid x)p(y \mid x)p(x)\log q(y \mid z)dxdydz \\
&= \int p(z \mid x)p(y,x)\log q(y \mid z)dxdydz,
\end{aligned} \tag{3.4}
$$

$$
\begin{aligned}
I(x;z) &= \int p(x,z)\log p(z \mid x)dxdz - \int p(z)\log p(z)dz \\
&\leq \int p(x,z)\log p(z \mid x)dxdz - \int p(z)\log q(z)dz \\
&= \int p(x)p(z \mid x)\log \frac{p(z \mid x)}{q(z)}dxdz \\
&= \int p(x,y)p(z \mid x)\log \frac{p(z \mid x)}{q(z)}dxdzdy,
\end{aligned} \tag{3.5}
$$

Bằng cách áp dụng giới hạn dưới của $I(y;z)$ và giới hạn trên của $I(x;z)$, hàm Lagrange trong phương trình (3.1) được xấp xỉ như

$$
\begin{aligned}
L_{IB} &= I(y;z) - \beta I(x;z) \\
&\approx \int p(z \mid x)p(y,x)\log q(y \mid z)dxdydz - \beta \int p(z \mid x)p(x,y)KL\Big(p(z \mid x)||q(z)\Big)dxdydz \\
&= \mathbb{E}_{(x,y)\sim p(x,y),z\sim p(z|x)}\Big[\log q(y \mid z) - \beta KL\Big(p(z \mid x)||q(z)\Big)\Big],
\end{aligned} \tag{3.6}
$$

Trong ứng dụng phát hiện tàu biển, thuật ngữ $q(y \mid z)$ được mô hình hóa bởi một bộ phân loại; và $\log q(y \mid z)$ là một classification loss $L_{cls}(\hat{y}_{cls}, y_{cls})$. Ngoài ra, đặc trưng ẩn z được lấy mẫu từ thủ thuật đo tham số lại $g(\varepsilon, x)$ nơi $\varepsilon \sim p(\varepsilon) = \mathcal{N}(0, I)$. Do đó,$z$ được ước tính bởi phương trình (3.7).

$$
z = \mu + \varepsilon * \sigma, \tag{3.7}
$$

Sử dụng phương trình (3.7), giá trị $p(z \mid x)$ được ước tính bởi phương trình (3.8). Giả định $q(z) = \mathcal{N}(0, I)$, giá trị $KL\Big(p(z \mid x)||q(z)\Big)$ sẽ được ước tính bởi phương trình (3.9). Ngoài ra, giá trị $KL\Big(p(z \mid x)||q(z)\Big)$ đóng vai trò như một thành phần điều chuẩn trong hàm mục tiêu VIB để lựa chọn đặc trưng, được ký hiệu là $L_{KL}(\mu, \sigma)$ trong phương trình (3.10). Do đó, tham số β được thay thế bởi tham số

Bảng 3.1: Ký hiệu toán học

Ký hiệu	Mô Tả
x, z, y	Đầu vào, đặc trưng được rút trích, và đầu ra của mạng.
i	Chỉ tiêu mức độ tỷ lệ (scale).
j	Chỉ số vị trí trên bản đồ đặc trưng.
d	Kích thước của vector ẩn trong mô-đun VIB
$\mu^i \in \mathbb{R}^{dH^iW^i}, \sigma^i \in \mathbb{R}^{dH^iW^i}$	đặc trưng ẩn và phương sai tương ứng của nó tại tỷ lệ thứ i^{th} .
$n_{detected}$	là tổng số lượng đối tượng được mô hình phát hiện.
$\mu_j \in \mathbb{R}^d, \sigma_j \in \mathbb{R}^d$	đặc trưng và phương sai tương ứng của nó tại vị trí j^{th} trong bản đồ
y_{cls}, $\hat{y}_{cls}$	Nhãn và kết quả dự đoán của bộ phân loại.
y_{reg}, $\hat{y}_{reg}$	Nhãn và kết quả dự đoán của nhánh dựa đoán đường biên.
y_{object} , $\hat{y}_{object}$	Nhãn và kết quả dự đoán của nhánh phát hiện đối tượng.

α_{KL}. Lưu ý rằng $L_{KL}(\mu, \sigma)$ trong phương trình (3.9) được áp dụng ở mọi mức tỷ lệ với hàm mục tiêu phân loại, hàm mục tiêu đường bao và hàm mục tiêu phát hiện đối tượng.

$$p(z \mid x) = \mathcal{N}(\mu, \sigma^2), \tag{3.8}$$

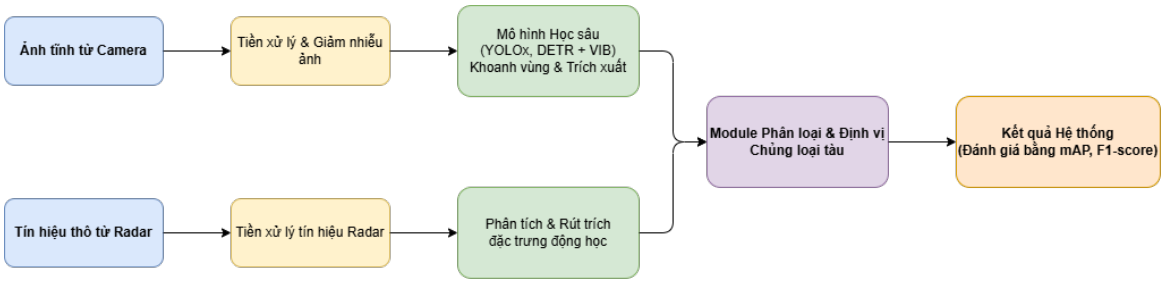
$$L_{KL}(\mu, \sigma) = KL\Big(p(z \mid x) || q(z)\Big) = \sum_{j=1}^{WH} (\mu_j^2 + \sigma_j^2 - 2log(\sigma_j) - 1), \tag{3.9}$$

Để kiểm chứng tính hiệu quả của kiến trúc mạng VIB-YOLOX vừa đề xuất, đặc biệt là khả năng hoạt động trên các mạng xương sống khác nhau, các thí nghiệm đánh giá chi tiết đã được thực hiện và phân tích trong mục dưới đây

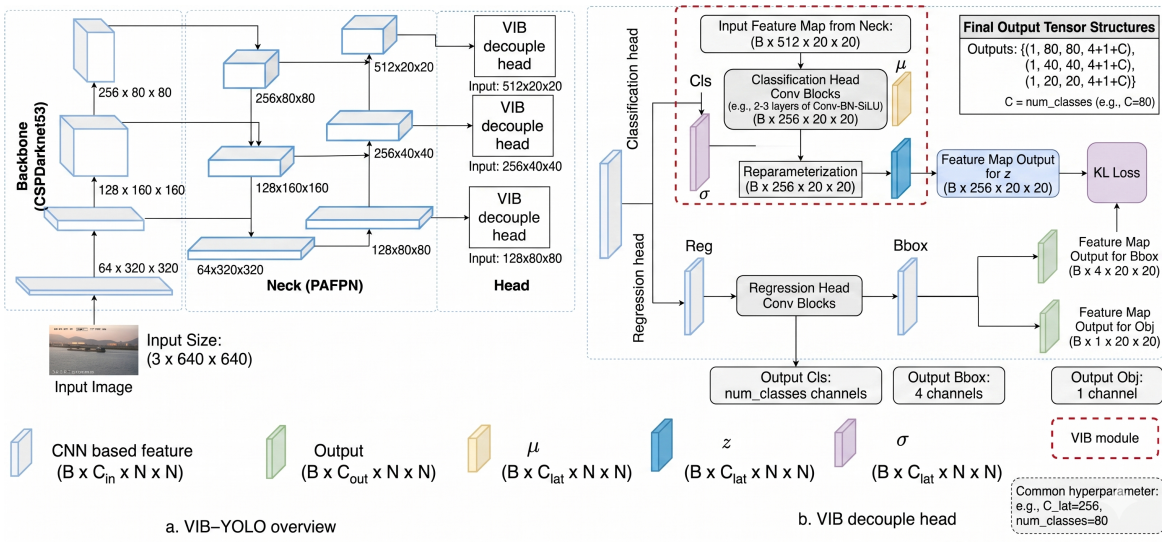
3.3.2 Tổng quan Phương pháp đề xuất

Bảng. 3.1 tóm tắt các ký hiệu toán học trong phương pháp đề xuất, và Hình. 3.8 giới thiệu khái niệm về phương pháp đề xuất. Dựa trên một bộ rút trích đặc trưng, các đặc trưng ở các tỷ lệ khác nhau được trích xuất. Ở đây, chúng tôi sử dụng mạng Darknet53 [2] để trích xuất các đặc trưng ở nhiều tỷ lệ. PAFPN [60] đóng vai trò là khối kết nối các đặc trưng này với nhánh dự báo. nhánh dự báo bao gồm đầu phân loại và đầu hồi quy. Trong khi phần đầu phân loại nhằm mục đích phân loại một loại tàu, thì phần đầu hồi quy ước tính một hộp giới hạn tương đối và khả năng đối tượng cho từng ô. Trên nhánh phân loại, chúng tôi sử dụng $1 * 1$ kernels để trích xuất $\mu_j \in \mathbb{R}^d$ và $\sigma_j \in \mathbb{R}^d$ tại j^{th} vị trí trên bản đồ đặc trưng. Sử dụng các kernel này, tensors $\mu \in \mathbb{R}^{dxHxW}$ và $\sigma \in \mathbb{R}^{dxHxW}$ thu được. Các ten-xơ này được

sử dụng để ước tính VIB loss [65]; Ngoài ra, một quá trình tái tham số hóa lấy mẫu một đặc trưng mới $z_j \in \mathbb{R}^d$ cho vị trí j^{th} . Một bộ phân loại lấy đặc trưng $z \in \mathbb{R}^{dHW}$ và dự đoán loại tàu $\hat{y}_{cls} \in \mathbb{R}^{KHW}$. Chi tiết của mô-đun VIB được mô tả trong bảng 3.2. Kích thước kernel $(1,1)$ có nghĩa là $Encoder_\mu$ trích xuất đặc trưng trao đổi chéo nhưng không thay đổi kích thước của bản đồ đặc trưng. Nó cho phép chúng tôi sử dụng lại đầu phân loại ban đầu.



Hình 3.7: Sơ đồ khối tổng thể hệ thống nhận dạng và phân loại tàu biển



Hình 3.8: kiến trúc mạng: (a) tổng quan phương pháp phát hiện đối tượng dựa trên VIB; (b) nhánh phân loại dựa trên VIB được đề xuất.

Để huấn luyện bộ phát hiện đối tượng, hành mục tiêu về phân loại $(L_{cls}(\hat{y}_{cls}, y_{cls}))$, hàm mục tiêu về đường bao $(L_{box}(\hat{y}_{box}, y_{box}))$ và hàm mục tiêu về đối tượng $(L_{obj}(\hat{y}_{obj}, y_{obj}))$ là những hàm mục tiêu thường xuyên được sử dụng. Nghiên cứu sinh cũng giới thiệu một hàm mục tiêu $L_{KL}(\mu, \sigma)$ để lựa chọn đặc trưng bằng cách sử dụng kỹ thuật information bottleneck. Phương trình 3.10 mô tả hàm mục tiêu tổng quát để huấn luyện bộ phát hiện đối tượng. Ở đây, các tham số α_{box}, α_{obj}, α_{cls}, α_{KL} mô tả sự đóng góp của mỗi hàm mục tiêu vào trong quá trình huấn luyện. Tiếp theo, phần 3.3.1 sẽ mô tả hàm mục tiêu để lựa chọn các đặc trưng, và các khối của mô hình (bộ rút trích đặc trưng và phần kết nối) sẽ được thảo luận ở phần 3.3.3.

Bảng 3.2: Chi tiết mạng. Ở đây, $Encoder_{\mu}$ là bộ mã hóa để trích xuất trị trung bình μ, $Encoder_{\sigma}$ là bộ mã hóa để trích xuất độ lệch chuẩn σ, $i$là chỉ số của cấp độ quy mô, và C^i là số lượng kênh trong đầu vào của cấp độ i^{th}.

Khối	Lớp	Thông số
$Encoder^i_{\mu}$	nn.conv	Size=(1,1), In = C^i, Out = C^i
$Encoder^i_{\sigma}$	nn.conv	Size=(1,1), In = C^i, Out = C^i
Re-parametrize	$z = \mu + \varepsilon * \sigma$	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$

$$L(\hat{y}, y) = \alpha_{box} L_{box}(\hat{y}_{box}, y_{box}) + \alpha_{obj} L_{obj}(\hat{y}_{obj}, y_{obj}) + \alpha_{cls} L_{cls}(\hat{y}_{cls}, y_{cls}) + \alpha_{KL} L_{KL}(\mu, \sigma), \quad (3.10)$$

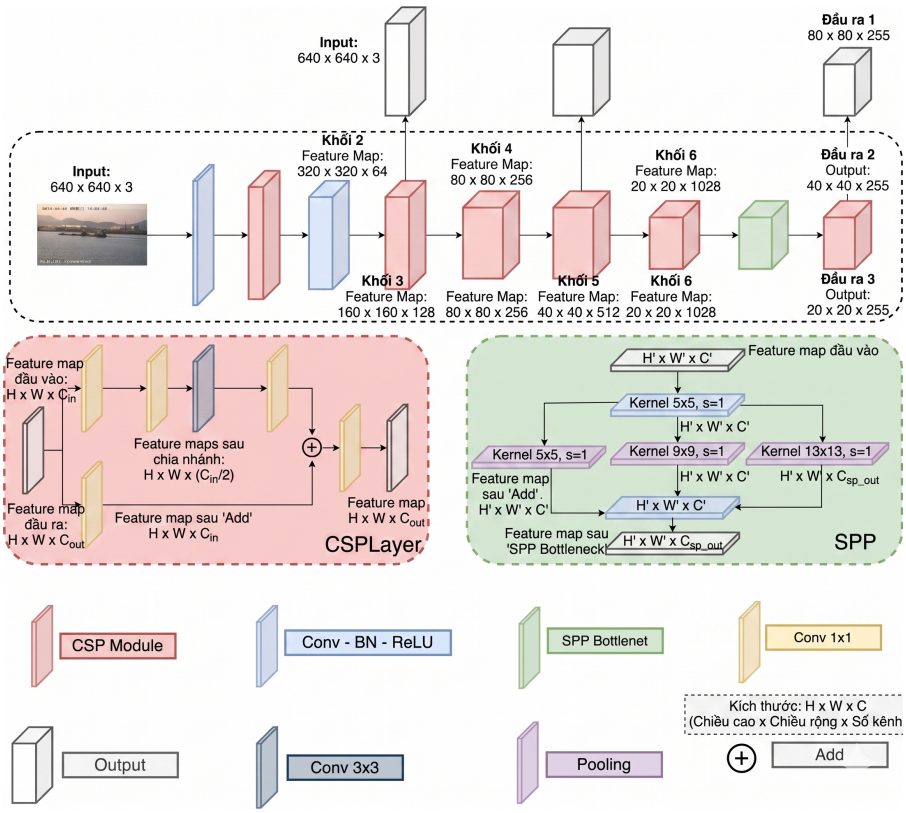
3.3.3 Khối trích xuất đặc trưng và khối kết nối

Phương pháp được đề xuất có thể được tích hợp với bất kỳ khối rút trích đặc trưng nào. Tuy nhiên, cần phải sửa đổi lớp trung gian (neck) để phù hợp với khối rút trích đặc trưng đã chọn và phần bộ phân loại đối tượng. Chúng tôi đã thử nghiệm một số bộ rút trích đặc trưng trong phần 3.5.2 và chỉ ra rằng bộ rút trích đặc trưng Darknet và khối kết hợp PAFPN cho kết quả tốt hơn các sự kết hợp khác.

Chi tiết về Darknet và PAFPN tương ứng trong Hình. 3.9 và Hình. 3.10. Ở đây, đường trục Darknet đã sử dụng CSPLayer để trích xuất các đặc trưng tại 2^{nd}, 3^{rd}, và 4^{th} CSPlayer được sử dụng trong phần kết nối PAFPN. Cuối cùng, các đặc trưng đầu ra được sử dụng với các nhánh phát hiện đối tượng và nhánh phát hiện đường bao ở các tỷ lệ khác nhau.

3.4 Bộ dữ liệu và ngữ cảnh thử nghiệm

Để đánh giá chất lượng cho phương pháp đề nghị, nghiên cứu này sử dụng bộ dữ liệu SeaShips [62]. Bộ dữ liệu SeaShips được xây dựng dựa trên các hình ảnh được chụp bởi một hệ thống giám sát video tại hiện trường được triển khai xung quanh đảo Hengqin, thành phố Châu Hải, Trung Quốc. Mỗi camera ghi lại cảnh nhiều tàu ra vào cảng từ 6 giờ sáng đến 8 giờ tối mỗi ngày. Theo tình hình thực tế của môi trường biển ở đảo Hengqin, bộ dữ liệu này nhóm tất cả các tàu thành sáu loại cụ thể bao gồm: tàu chở quặng, tàu chở hàng rời, tàu chở hàng tổng hợp, tàu container, tàu đánh cá và tàu chở khách và dán nhãn cho chúng. Bộ dữ liệu [62] gồm 60 clip. Mỗi clip ít nhất 60 giây. Cứ sau hai giây, một hình ảnh được chụp. Tập dữ liệu [62] có 31455 hình ảnh mỗi hình ảnh bao gồm nhiều tàu khác

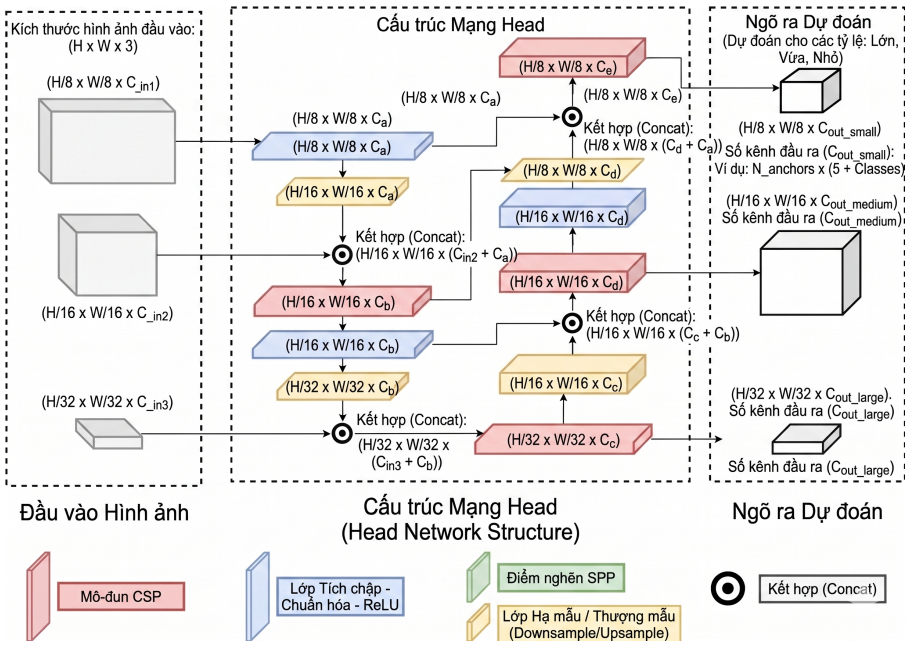


Hình 3.9: Cấu trúc Darknet

nhau. Tuy nhiên, chỉ có 7000 hình ảnh được công bố để nghiên cứu tại [66, 56, 47].

Tổng quan tài liệu cho thấy nhiều công trình nghiên cứu sử dụng 80% dữ liệu đã công bố để huấn luyện/xác nhận và 20% để thử nghiệm. Do đó, chúng tôi chọn D_1^{Train} bao gồm 5600 hình ảnh để huấn luyện và D_1^{Test} bao gồm 1400 hình ảnh để thử nghiệm. Ngoài ra, những công trình gần đây [61, 46] cũng sử dụng các ngữ cảnh thách thức hơn, trong đó 50% dữ liệu là tập huấn luyện và phần còn lại là tập kiểm tra. Chúng tôi cũng tuân theo ngữ cảnh này để chuẩn bị D_2^{Train} và D_2^{Test} để so sánh. Để đánh giá hiệu suất trên một tập dữ liệu rất nhỏ, chúng tôi chọn ngẫu nhiên một số tập dữ liệu con S_1, S_2, S_3 bao gồm 30%, 70% và 100% mẫu từ D_2^{Train} để huấn luyện về các thí nghiệm sau này.

Các thí nghiệm sau đây sử dụng kỹ thuật SGD, tốc độ học tập = 0,01, tỷ lệ suy giảm tốc độ học = 0,001, n_epoch=200 và batch_size=8. Ngoài ra, các hàm mục tiêu cho VIB sử dụng toán tử trung bình trên mỗi batch_size và toán tử tổng trên nhánh phân loại. Chỉ số mAP được sử dụng để chọn mô hình tốt nhất trong suốt quá trình huấn luyện. Các hệ số trong hàm mục tiêu của phương trình 3.10 được lựa chọn như sau $\alpha_{box} = 10,0; \alpha_{obj=}1,0; \alpha_{cls} = 1,0$ và $\alpha_{KL} = 0,125$.

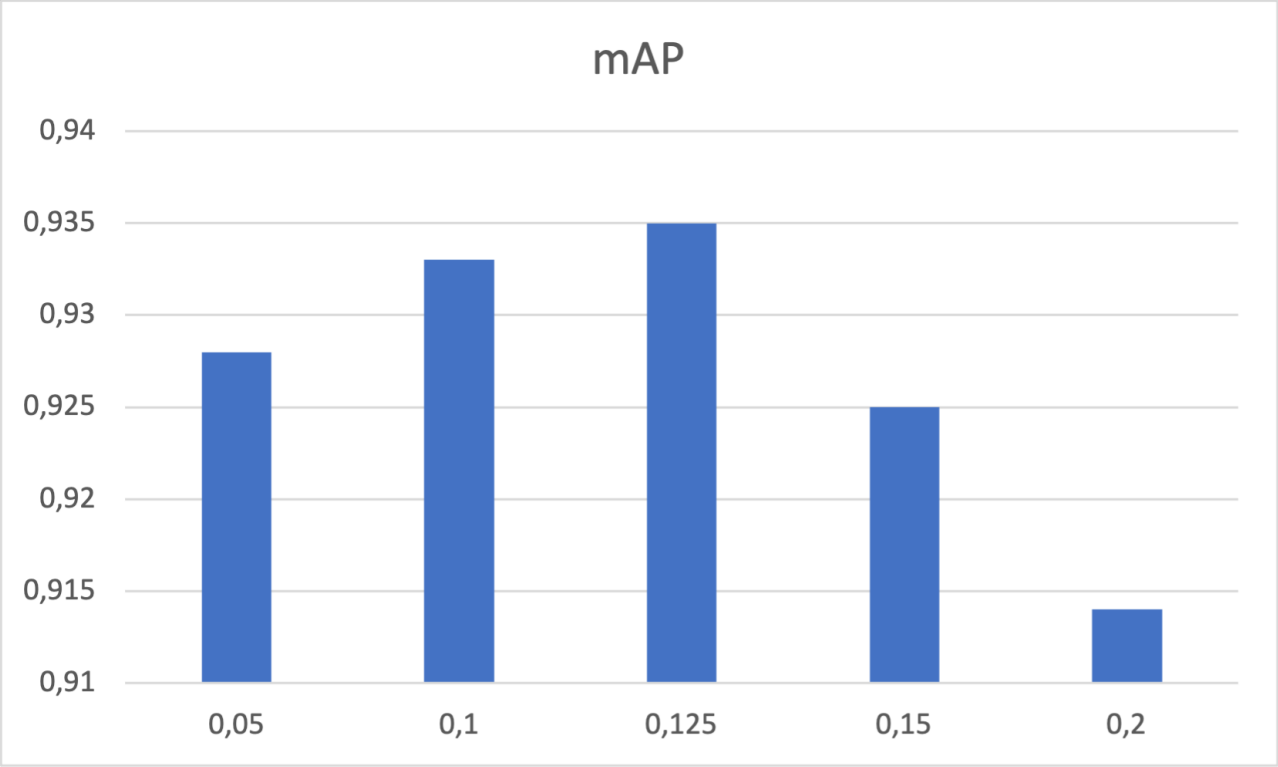


Hình 3.10: Cấu trúc PAFPN

3.5 Kết quả thực nghiệm với mạng YOLOX

3.5.1 Chọn siêu tham số α_{KL}

Phần này thảo luận về việc chọn một siêu tham số phù hợp α_{KL} trong phương trình 3.10. Tham số này giống với β trong phương trình 3.6. Một α_{KL} nhỏ có thể không giúp tìm hiểu các đặc trưng tốt hơn, trong khi một α_{KL} lớn có thể tập trung quá nhiều vào việc học đặc trưng và quên nhiệm vụ chính. Trong thí nghiệm này, các S_2 tập dữ liệu được sử dụng để huấn luyện, và D_2^{Test} tập dữ liệu được sử dụng để kiểm tra. Số liệu mAP trên sáu lớp được sử dụng để so sánh. Hình 3.11 cho thấy cách thức mà hàm mục tiêu KL ảnh hưởng đến kết quả. Khi không sử dụng kỹ thuật lựa chọn đặc trưng, giá trị mAP chỉ đạt được 0,923; khi α_{KL} bằng 0,05, giá trị mAP tăng lên 0,928. Khi α_{KL} nhận giá trị cao hơn, mAP nhận được ngày càng cao. Tuy nhiên, khi $\alpha_{KL} = 0,15$, mAP bắt đầu giảm; và nếu $\alpha_{KL} = 0,2$ mAP là 0,914. Giá trị mAP dựa trên $\alpha_{KL} = 0,2$ nhỏ hơn so với khi không sử dụng hàm mục tiêu VIB để lựa chọn đặc trưng. Hiện tượng này là do hàm mục tiêu lựa chọn đặc trưng làm sẽ làm giảm các đặc trưng được chọn cho tác vụ chính. Khi số lượng đặc trưng bị giảm quá nhiều, bộ phân loại phân loại có thể không có đủ thông tin cho nhiệm vụ phân loại. Trong các thí nghiệm tiếp theo, chúng tôi chọn $\alpha_{KL} = 0,125$ cho phương pháp đề xuất của chúng tôi.



Hình 3.11: Giá trị mAP ứng với các tham số α_{VIB} khác nhau. Trục x có nghĩa mô tả tham số α_{KL}, và trục y mô tả độ chính xác trung bình trung bình trên tất cả các lớp.

3.5.2 Ảnh hưởng của việc lựa chọn các bộ rút trích đặc trưng

Phần này thảo luận về cách thức hoạt động của phương pháp được đề xuất với các bộ rút trích đặc trưng khác nhau. ResNet, MobileNetv2 và DarkNet được sử dụng làm các bộ rút trích đặc trưng để so sánh. Bộ kết hợp được điều chỉnh để đáp ứng đầu ra của các bộ rút trích đặc trưng này. Độ chính xác trung bình (AP) cho sáu lớp được hiển thị trong Bảng 3.4. Trong thử nghiệm này, S_2 đóng vai trò là tập dữ liệu huấn luyện. Kết quả cho thấy DarkNet là bộ rút trích đặc trưng tốt nhất trong số các mô hình được huấn luyện này. Điều này là hợp lý vì DarkNet đã được công nhận là mạng cơ sở tốt nhất trong các phiên bản YOLO. Ngoài ra, mức độ cải thiện do VIB tạo ra trên ResNet [67] là 5,9 %. Điều đó có nghĩa là hàm mục tiêu VIB có thể hỗ trợ rất nhiều cho một số bộ rút trích đặc trưng cụ thể. Bên cạnh các chỉ số về độ chính xác, phân tích chi phí tính toán cũng được thực hiện để đánh giá khả năng ứng dụng thực tế. Bảng 3.4 trình bày tốc độ khung hình trên giây (FPS) tương ứng với từng cấu hình. Việc tích hợp mô-đun VIB đòi hỏi thêm một lượng nhỏ tính toán, dẫn đến tốc độ suy luận giảm nhẹ (khoảng 3-4 FPS tùy mạng). Tuy nhiên, mức giảm này là hoàn toàn chấp nhận được khi đối chiếu với sự gia tăng đáng kể của mAP.Sự đánh đổi (trade-off) giữa độ chính xác và tốc độ suy luận thể hiện rõ rệt qua các mạng xương sống. MobileNetv2 cung cấp tốc độ xử lý nhanh nhất (81 FPS khi có VIB),

biến nó thành lựa chọn lý tưởng cho các hệ thống nhúng, thiết bị biên (edge devices) hoặc drone có tài nguyên phần cứng hạn chế, chấp nhận mức mAP thấp hơn một chút. Ngược lại, DarkNet mang lại độ chính xác cao nhất (mAP đạt 0,935) nhưng tốc độ suy luận dừng ở mức 42 FPS, cấu hình này đặc biệt phù hợp cho các trạm quan trắc trên bờ, nơi có thể trang bị các máy chủ GPU mạnh mẽ. ResNet-18 đứng ở vị trí cân bằng với 62 FPS, dung hòa tốt giữa tốc độ và độ chính xác cho các hệ thống giám sát tầm trung. Tùy thuộc vào yêu cầu của hệ thống triển khai thực tế mà người dùng có thể linh hoạt chọn kiến trúc mạng tương ứng.

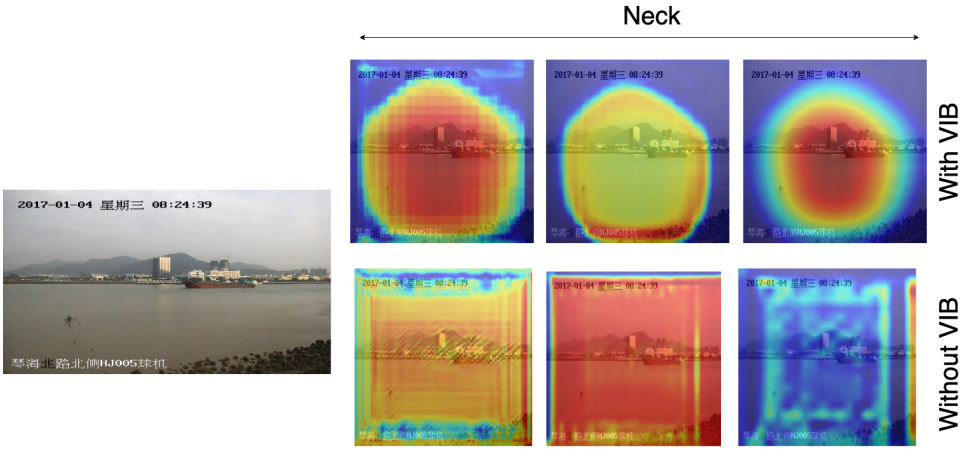
Bảng 3.3: So sánh mAP có và không có VIB trên các bộ rút trích đặc trưng khác nhau.

mạng xương sống	**Với VIB**	**fishing boat**	**container ship**	**ore carrier**	**bulk cargo carrier**	**passenger ship**	**general cargo ship**	**mAP**
ResNet -18	có	0,837	0,978	0,861	0,886	0,741	0,931	**0,873**
	không có	0,817	0,961	0,824	0,735	0,706	0,841	0,814
DarkNet	có	0,922	0,980	0,935	0,953	0,873	0,946	**0,935**
	không có	0,903	0,970	0,932	0,928	0,862	0,945	0,923
Mobile Netv2	có	0,895	0,975	0,895	0,880	0,794	0,908	**0,891**
	không có	0,872	0,965	0,918	0,902	0,772	0,914	0,890

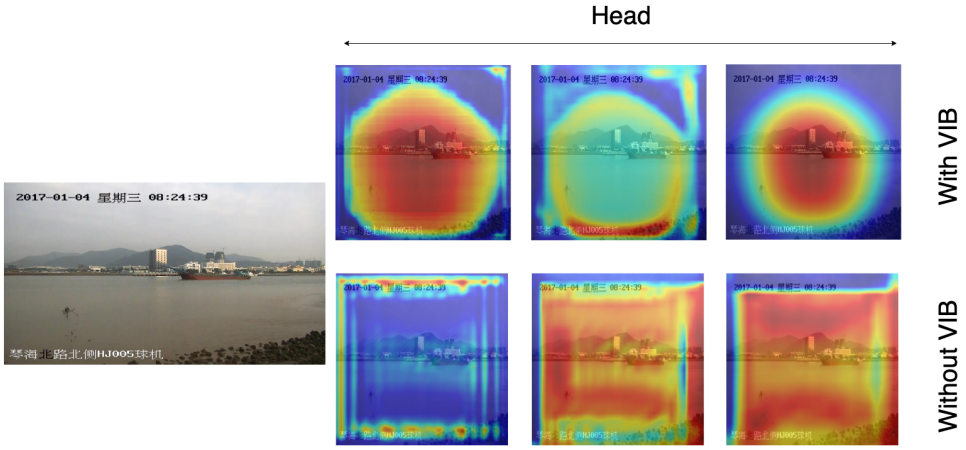
Bảng 3.4: So sánh mAP và tốc độ suy luận (FPS) có và không có VIB trên các bộ rút trích đặc trưng khác nhau.

Mạng xương sống	**Với VIB**	**fishing boat**	**container ship**	**ore carrier**	**bulk cargo carrier**	**passenger ship**	**general cargo ship**	**mAP**	**FPS**
ResNet 18	có	0,837	0,978	0,861	0,886	0,741	0,931	**0,873**	62
	không có	0,817	0,961	0,824	0,735	0,706	0,841	0,814	65
DarkNet	có	0,922	0,980	0,935	0,953	0,873	0,946	**0,935**	42
	không có	0,903	0,970	0,932	0,928	0,862	0,945	0,923	45
Mobile Netv2	có	0,895	0,975	0,895	0,880	0,794	0,908	**0,891**	81
	không có	0,872	0,965	0,918	0,902	0,772	0,914	0,890	85

Để giải thích rõ ràng lợi ích của phương pháp được đề xuất đối với việc học đặc trưng, chúng tôi so sánh đặc trưng được học theo phương pháp của chúng tôi (với VIB) và phương pháp cơ sở (không có



Hình 3.12: Bản đồ đặc trưng trên lớp trung gian của bộ phân loại



Hình 3.13: Bản đồ đặc trưng trên tầng đầu ra của bộ phân loại

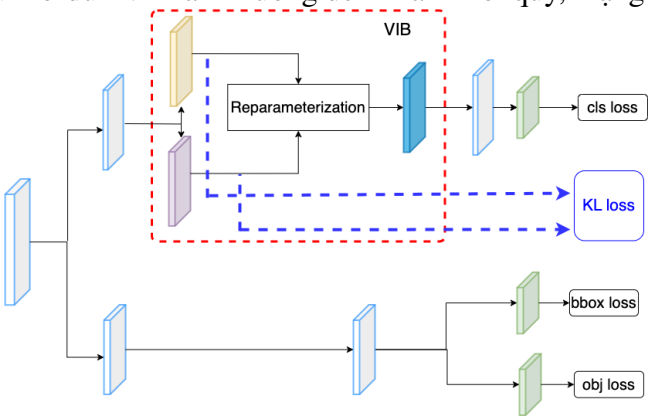
VIB) bằng cách sử dụng bộ dữ liệu S_3. Các đặc trưng trước khi lớp trung gian và tầng đầu cuối cùng được trích xuất và trực quan hóa. Chúng tôi chọn 20 bản đồ đặc trưng có điểm phản hồi cao nhất cho mỗi cấp độ tỷ lệ. Gọi j là một điểm ảnh trên bản đồ đặc trưng F có kích thước (W,H); mức điểm của bản đồ đặc trưng là $\frac{1}{WH}\sum_{j=1}^{WH} F_j$. Các bản đồ đặc trưng này được cộng dồn với nhau để tạo thành một bản đồ duy nhất thể hiện sự ảnh hưởng của một điểm ảnh. Kết quả có thể đại diện cho các điểm quan trọng trên ảnh đầu vào.

Hình 3.12 và Hình 3.13 hiển thị bản đồ nhiệt tương ứng với hình ảnh đầu vào. Hàng đầu tiên thể hiện bản đồ đặc trưng khi có sử dụng VIB; hàng thứ hai đại diện cho những bản đồ đặc trưng không sử dụng VIB. Bởi vì các đặc trưng được trích xuất ở ba cấp độ tỷ lệ, ba phản hồi được cung cấp cho các mô-đun phân loại. Kết quả cho thấy, dưới sự định hướng của hàm mất mát VIB, mạng nơ-ron có khả năng trích xuất các đặc trưng tập trung vào đối tượng (loại bỏ được các nhiễu nền). Không có VIB, phản hồi thích phân phối đồng đều. Với VIB, các phản hồi của đặc trưng tập trung xung quanh đối

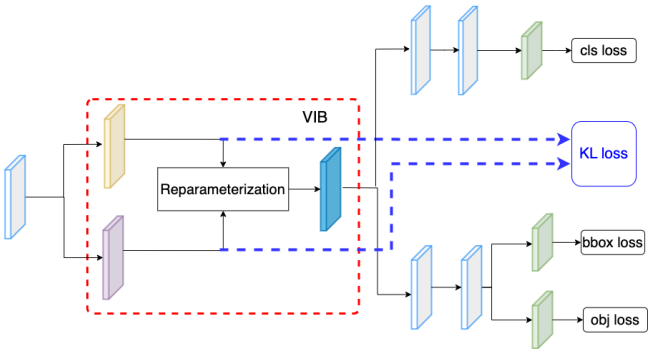
tượng chứ không phải tất cả các pixel. Hiện tượng này cũng lặp lại ở lớp trung gian. Điều đó có nghĩa là tổn thất VIB có thể được lan truyền ngược về các khối kết hợp và học được các đặc trưng tốt hơn.

3.5.3 Ảnh hưởng của vị trí của khối VIB trong mô hình phát hiện đối tượng

Trong phương pháp đề xuất, Nghiên cứu sinh đã chèn khối VIB vào vị trí bắt đầu của nhánh phân loại. Cần lưu ý rằng để đánh giá hiệu quả của việc chèn khối VIB, thực nghiệm trong trường hợp này được tiến hành với bộ rút trích đặc trưng là mạng DarkNet. Tuy nhiên, mô-đun VIB có thể được chèn vào bất kỳ vị trí nào của kiến trúc mạng. Do đó, trong phần này, chúng tôi đã thử một số thiết lập để đánh giá cách sử dụng VIB trong tác vụ phát hiện đối tượng. Trong YOLOX, nhánh phân loại có hai khối tích chập được sắp xếp nối tiếp. Phương pháp được đề xuất sẽ chèn khối VIB vào đầu nhánh phân loại, như trong Hình 3.8. Tuy nhiên, chúng ta cũng có thể thiết lập mô-đun VIB ở giữa nhánh phân loại như trong Hình 3.14 hoặc ở phần bắt đầu của cả hai nhánh phân loại và phát hiện đường bao như trong Hình 3.15. Ở đây, tập dữ liệu S_2 được sử dụng để huấn luyện như trong Phần 3.5.2. Kết quả trong Bảng 3.5 cho thấy mô-đun VIB chỉ phù hợp để chèn vào nhánh phân loại. Nếu mô-đun VIB ảnh hưởng đến nhánh hồi quy, mạng không thể hội tụ. caption



Hình 3.14: VIB ở phần giữa của đầu phân loại

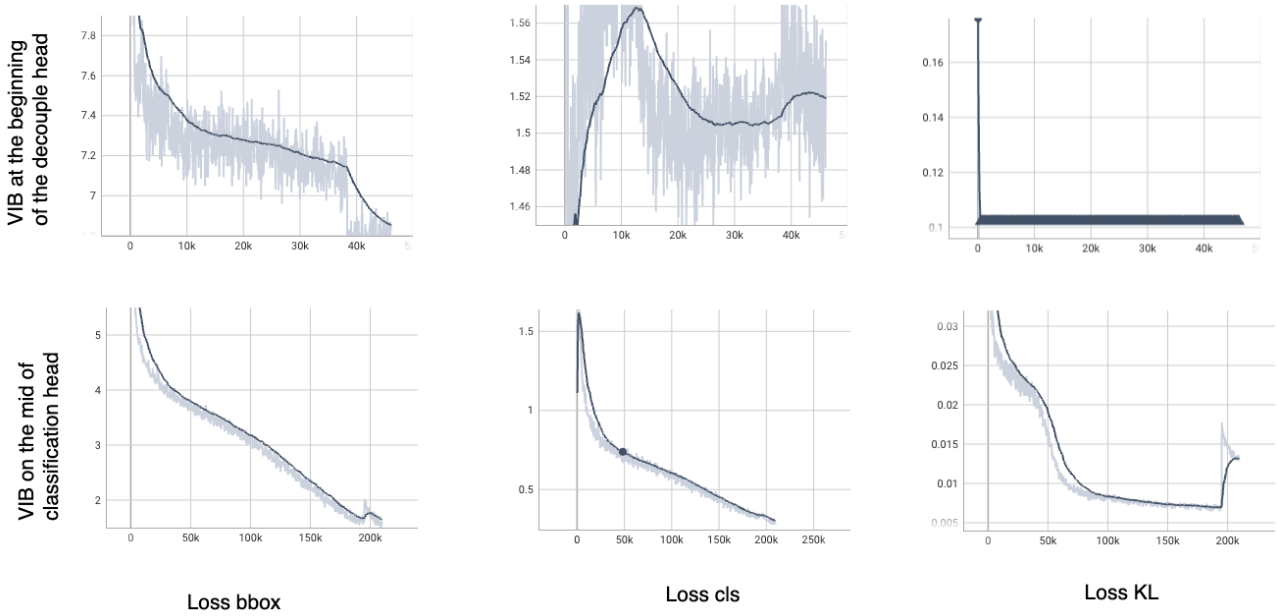


Hình 3.15: VIB ở phần bắt đầu của đầu tách

Bảng 3.5: Đánh giá hiệu quả dựa trên vị trí chèn mô-dun VIB.

Phương pháp	**Chỉ số**	**fishing boat**	**container ship**	**ore car-rier**	**bulk cargo carrier**	**passenger ship**	**general cargo ship**	mAP
Giữa cls đầu ra	$n_{detected}$	1863	584	2012	1653	354	1129	
	recall	**0,940**	**0,984**	0,962	**0,971**	0,891	0,964	
	AP	**0,922**	**0,980**	**0,935**	**0,953**	0,873	**0,946**	**0,935**
Sau lớp trung gian	$n_{detected}$	-	-	-	-	-	-	
	recall	-	-	-	-	-	-	
	AP	-	-	-	-	-	-	

Để giải thích hiện tượng này, các hàm mục tiêu L_{box}, L_{cls}, và L_{KL} trong quá trình huấn luyện được hiển thị trong Hình. 3.16.



Hình 3.16: Giá trị hàm mục tiêu trong một quá trình huấn luyện. Dòng in đậm là các giá trị được làm trơn qua các lần lặp. Đường nhạt hơn là giá trị thực tế trong một lần lặp.

mô hình có thể hội tụ thuận lợi nếu mô-đun VIB nằm trên nhánh phân loại (hàng thứ hai của Hình. 3.16). L_{box} nhanh chóng giảm xuống đến phạm vi $[4-5]$ chỉ với 3000 lặp lại. Thành công của nhánh dự đoán đường biên là một yêu cầu quan trọng để huấn luyện nhánh phân loại. Khi bắt đầu quá trình huấn luyện, L_{cls} tăng khi L_{box} lớn; sau đó, nó giảm dần khi L_{box} nhỏ hơn. L_{KL} sẽ đóng góp sau trong quá trình huấn luyện vì đây là một hàm mục tiêu phụ chứ không phải là nhiệm vụ chính cần giải quyết.

mô hình không thể hội tụ nếu mô-đun VIB ở bắt đầu của cả hai nhánh (hàng đầu tiên của Hình. 3.16).

L_{box} giảm nhưng vẫn cao hơn 7 sau lần lặp 40000. Khi nhánh dự đoán đường bao không hoạt động thành công, nhánh phân loại không thể huấn luyện được. Ta thấy, L_{cls} tăng và giảm trong quá trình huấn luyện. Điều đó là hợp lý vì không thể huấn luyện nhánh phân loại nếu L_{box} vẫn còn lớn. Hiện tượng này cho thấy việc hàm mất mát KL và quá trình tái lấy mẫu trên miền đặc trưng làm cho quá trình hồi quy trở nên khó khăn hơn. Do đó, không thể huấn luyện được nhánh phân loại và mô hình không thể hội tụ.

3.5.4 So sánh với các phương án tốt nhất trong phát hiện tàu biển

Phần này so sánh phương pháp được đề xuất của chúng tôi với các phương pháp tốt nhất đã được công bố trên chỉ số mAP. Nhiều thiết lập khác nhau đã được sử dụng dựa trên các bài báo trước đó. Dựa trên 70000 hình ảnh được xuất bản từ tập dữ liệu SeaShip [62]; Zhang_2022 [59], và Zhang_2021 [53] đã sử dụng 90% dữ liệu để huấn luyện và xác thực mô hình; 10% dữ liệu còn lại là tập dữ liệu thử nghiệm. Liu_2020 [48], Liu_2022 [45], Han_2021[55], và Light_SDNet [56] sử dụng 80% dữ liệu để huấn luyện và xác thực tập dữ liệu, 20% còn lại là tập dữ liệu thử nghiệm. Để so sánh với các nghiên cứu này, chúng tôi sử dụng D_1^{Train} để huấn luyện và D_1^{Test} để kiểm tra. Kết quả trong bảng 3.6 cho thấy phương pháp của chúng tôi tốt hơn các phương pháp khác về độ chính xác trung bình. Có hai lý do cho độ cải thiện này. Đầu tiên, phương pháp của chúng tôi dựa trên phương pháp YOLOX, là phương pháp tốt nhất gần đây để phát hiện đối tượng. Light_SDNet dựa trên YOLOv5 và kết quả của nó cũng đầy hứa hẹn. Liu_2020 [48], Liu_2022 [45] dựa trên các phiên bản cũ hơn trong nhóm các phiên bản YOLO; do đó hiệu suất nhỏ hơn các phương pháp tốt nhất đã công bố như Light_SDNet và phương pháp được đề xuất. Yếu tố thứ hai giúp cải thiện kết quả là việc tái lấy mẫu các đặc trưng trong qúa trình huấn luyện. Việc tái lấy mẫu sẽ cộng thêm nhiễu vào quá trình huấn luyện. Nó cho phép mô hình có thể hoạt động tốt ngay cả khi có nhiễu từ môi trường. Để so sánh, Light_SDNet [56] cũng thêm nhiều yếu tốt như sương mù và mưa vào ảnh gốc; nhờ đó mô hình phát hiện đối tượng nhận được kết quả rất tốt (mAP=0.988%). Sự khác biệt chính giữa phương pháp của chúng tôi và phương pháp Light_SDNet [56] là cách chúng tôi thêm nhiễu vào dữ liệu huấn luyện. Trong khi Light_SDNet [56] thêm nhiễu vào ảnh gốc, phương pháp đề xuất đã thêm nhiễu vào không gian đặc trưng. Mặc dù không phải là thành phần cốt lõi của kiến trúc mạng, hàm mất mát VIB (Variational Information Bottleneck Loss) vẫn giữ vai trò quan trọng trong quá trình huấn luyện. Cụ thể, VIB Loss thúc đẩy mô hình học các đặc trưng giàu thông tin liên quan trực tiếp đến đối tượng mục tiêu, đồng thời hạn chế các đặc trưng dư thừa hoặc nhiễu xuất phát từ nền. Phần phân tích đặc trưng được trình bày tại Mục

3,7.5 3.5.5 sẽ cung cấp minh chứng định tính và định lượng thông qua các bản đồ đặc trưng nhằm làm rõ hiệu quả của cơ chế này.

Ngoài ra, Biaohua_2022 [46] và Yani_2022 [61] đã sử dụng 50% hình ảnh đã xuất bản để huấn luyện và 50% còn lại để thử nghiệm. Do đó, chúng tôi sử dụng D_2^{Train} và D_2^{Test} để huấn luyện và kiểm tra tương ứng. Như thể hiện trong Bảng 3.6, mAP trên các tác phẩm trước đó lên tới 0,965%. Phương pháp đề xuất của chúng tôi có thể có hiệu suất tốt hơn đáng kể so với các công trình trước đây. Nó chứng minh lợi ích của phương pháp của chúng tôi khi số lượng mẫu huấn luyện giảm.

Bảng 3.6: So sánh hiệu suất của các phương pháp khác nhau. Các kết quả tốt nhất được **in đậm**.

Method	**Train+Val / Test (in %)**	**fishing boat**	**container ship**	**ore carrier**	**bulk cargo carrier**	**passenger ship**	**general cargo ship**	**mAP**
Zhang_2022 [59]	90/10	0,824	0,940	0,859	0,915	0,787	0,914	0,873
Zhang_2021 [53]	90/10	-	-	-	-	-	-	0,946
Liu_2020 [48]	80/20	-	-	-	-	-	-	0,908
Liu_2022 [45]	80/20	-	-	-	-	-	-	0,964
Han_2021[55]	80/20	-	-	-	-	-	-	0,906
Light_SDNet [56]	80/20	**0,986**	0,995	**0,989**	0,990	0,982	0,989	0,988
Proposed method	80/20	0,979	**1**	0,987	**0,994**	**0,994**	**0,993**	**0,991**
Yani_2022 (ESDT) [61]	50/50	-	-	-	-	-	-	0,593
Yani_2022 (DETR)[61]	50/50	-	-	-	-	-	-	0,965
Biaohua_2022 [46]	50/50	0,940	0,987	0,966	0,978	0,937	0,972	0,963
Proposed method	50/50	0,970	0,986	0,984	0,991	0,964	0,989	**0,980**

3.5.5 Thí nghiệm trên các tập dữ liệu nhỏ

Trong phần này, chúng tôi thảo luận về sự đóng góp của hàm mục tiêu VIB trên các bộ dữ liệu có quy mô khác nhau. Bộ dữ liệu S_1 (525 ảnh), S_2 (1225 ảnh), và S_3 (3500 ảnh) được sử dụng để huấn luyện. Bộ dữ liệu thử nghiệm D_2^{Test} bao gồm 3500 ảnh. Kết quả trong Bảng. 3.7 cho thấy hàm mục tiêu VIB

giúp cải thiện hiệu suất đáng kể trên các tập dữ liệu nhỏ. Nếu tập dữ liệu huấn luyện là S_1, mAP được cải thiện 3%. Khi số lượng mẫu huấn luyện tăng lên, độ cải thiện sẽ không còn nhiều như trước. Độ cải tiến trên mAP là 1,2% nếu S_2 được sử dụng để huấn luyện và nếu tập dữ liệu huấn luyện là S_3, thì các mAP từ cả hai ngữ cảnh (có và không có sử dụng hàm mục tiêu VIB) đều tương đối giống nhau. Hiện tượng này là hợp lý vì hàm mục tiêu không được giám sát có thể giúp tránh hiện tượng overfitting trên các tập dữ liệu nhỏ và việc tái lấy mẫu trên miền đặc trưng cho phép bộ phân loại trở nên ổn định.

Bảng 3.7: Hiệu suất trên các tập dữ liệu nhỏ. S_1 nghĩa là 30% mẫu huấn luyện, S_2 nghĩa là 70% mẫu huấn luyện, S_3 nghĩa là 100% mẫu huấn luyện từ D_2^{Train}. Các kết quả tốt nhất được **in đậm**.

Phương pháp	**Số liệu**	**fishing boat**	**container ship**	**ore carrier**	**bulk cargo carrier**	**passenger ship**	**general cargo ship**	mAP
	dets	2849	1272	5374	3581	774	3598	
S_1 + VIB	recall	0,878	0,941	**0,924**	**0,913**	**0,657**	**0,946**	
	AP	0,796	0,884	0,831	**0,765**	**0,524**	**0,794**	**0,766**
	dets	3201	884	4654	2851	696	2944	
S_1 no VIB	recall	**0,892**	0,920	0,923	0,876	0,637	0,940	
	AP	**0,805**	**0,890**	0,822	0,710	0,466	0,743	0,739
	dets	1863	584	2012	1653	354	1129	
S_2 + VIB	recall	**0,940**	**0,984**	0,962	**0,971**	0,891	0,964	
	AP	**0,922**	**0,980**	**0,935**	**0,953**	0,873	**0,946**	**0,935**
	dets	2324	674	2848	1923	570	1585	
S_2 no VIB	recall	0,936	0,975	0,964	**0,963**	**0,899**	**0,978**	
	AP	0,903	0,970	0,932	0,928	0,862	0,945	0,923
	dets	1544	466	1562	1341	297	948	
S_3 + VIB	recall	**0,978**	0,986	0,990	**0,995**	**0,972**	**0,993**	
	AP	**0,970**	0,986	0,984	**0,991**	**0,964**	**0,989**	**0,98**
	dets	1574	470	1561	1360	294	883	
S_3 no VIB	recall	0,965	**0,989**	**0,995**	0,993	0,960	**0,993**	
	AP	0,957	**0,988**	0,990	0,987	0,953	0,989	0,977

3.6 Kết quả thí nghiệm với DETR

Trong phần này, luận án trình bày các kết quả thực nghiệm đối với họ mô hình DETR. Tuy nhiên, để khắc phục nhược điểm về tốc độ hội tụ và tối ưu hóa khả năng trích xuất đặc trưng không gian (đặc biệt đối với các đối tượng tàu biển có kích thước đa dạng), kiến trúc Deformable DETR đã được lựa chọn làm mô hình thực nghiệm chính.

Do đó, cần làm rõ rằng toàn bộ các đánh giá trong mục này đều dựa trên biến thể Deformable DETR. Cụ thể, các cấu hình thử nghiệm và kết quả phân tích đặc trưng được trình bày tại Bảng 3,7, cũng như hiệu năng tổng thể của mô hình được tóm tắt tại Bảng 3,8, đều là kết quả thực thi trực tiếp từ mô hình Deformable DETR kết hợp với phương pháp đề xuất của luận án. Để đạt được các kết quả này, kiến trúc Encoder-Decoder và cơ chế Attention đa mức của Deformable DETR đã được tinh chỉnh chuyên sâu. Điểm cốt lõi trong phương pháp đề xuất là việc tích hợp kỹ thuật Nút thắt thông tin biến phân (VIB) nhằm triệt tiêu nhiễu bối cảnh biển, kết hợp cùng hàm mất mát phân đôi được tối ưu riêng cho đặc thù hình học thon dài của tàu thuyền.

3.6.1 Lựa chọn siêu tham số

Như đã thảo luận trong Phần 3, số lượng truy vấn có thể ảnh hưởng lớn đến số lượng đầu ra của một bộ phát hiện dựa trên DETR. Do đó, phần này tập trung vào việc lựa chọn tham số phù hợp nhất để kiểm soát mô hình. Chúng tôi tiến hành so sánh hiệu suất khi $n_{queries}$ nhận các giá trị lần lượt trong danh sách [300, 200, 100, 50]. Các tập D_2^{Train} và D_2^{Test} được chọn làm tập huấn luyện và tập kiểm thử để đảm bảo một thiết lập bài toán mang tính thử thách cao. Điều này có nghĩa là 50% dữ liệu được sử dụng cho quá trình kiểm thử.

Các kết quả thực nghiệm được trình bày chi tiết trong Bảng 3.8. Trong thư viện mmdetection, giá trị mặc định cho $n_{queries}$ là 300. Quan sát cho thấy việc sử dụng thiết lập mặc định này dẫn đến sự gia tăng đáng kể số lượng phát hiện. Ví dụ, số lượng phát hiện đối với lớp fishing boat là 204,449. Xu hướng này dẫn đến việc giảm độ chính xác trung bình (AP) xuống 0,798, trong khi độ tăng lên 0,913. Ngoài ra, số lượng phát hiện cao hơn đối với tàu cá có thể là do tần suất xuất hiện cao hơn của nhãn này trong tập dữ liệu.

Bằng cách giảm $n_{queries}$, sự thiên lệch trong các phát hiện được giảm thiểu đáng kể. Cụ thể, khi $n_{queries}$ được giảm xuống các mức 200, 100 và 50, số lượng phát hiện tàu cá giảm xuống tương ứng là 208817, 119774 và 96577. Hơn nữa, việc giảm tham số này giúp cân bằng số lượng phát hiện giữa các loại tàu khác nhau. Với $n_{queries} = 300$, số lượng phát hiện thấp nhất thuộc về tàu container (6300). Tuy nhiên, khi $n_{queries} = 50$, số lượng phát hiện thấp nhất tăng lên khoảng 18949 và sự phân bố giữa tàu container, tàu chở quặng và tàu khách là khá tương đồng. caption

Bên cạnh số lượng truy vấn, Learning Rate là một siêu tham số quan trọng khác ảnh hưởng trực tiếp đến hiệu suất hệ thống. Mặc định, Lr được thiết lập là 10^{-4}. Trong phiên bản sửa đổi này, chúng tôi

Bảng 3.8: Ảnh hưởng của số lượng truy vấn đối tượng(*queries*) tới kết quả phát hiện đối tượng.

Class	300 query			200 query			100 query			50 query		
	dets	**recall**	**AP**	**dets**	**recall**	**AP**	**dets**	**recall**	**AP**	**dets**	**recall**	**AP**
fishing boat	204449	0,913	0,798	208817	0,985	0,970	119774	**0,986**	**0,969**	96577	0,972	0,949
container ship	6300	0,989	0,894	11460	0,995	0,995	31083	0,995	0,989	18949	**0,998**	**0,997**
ore carrier	53520	0,987	0,923	49362	**0,997**	**0,992**	62122	0,996	0,989	18949	0,993	0,987
bulk cargo carrier	45541	0,985	0,912	47754	0,996	**0,992**	46490	**0,997**	0,986	76230	0,996	**0,989**
passenger ship	22498	0,968	0,610	16199	0,970	**0,957**	14866	0,968	0,936	20043	**0,972**	0,926
general cargo ship	17692	0,987	0,930	16408	0,995	**0,992**	75665	0,995	0,991	98537	**0,996**	**0,990**
mAP			0,844			**0,982**			0,977			0,973

thử nghiệm điều chỉnh Lr thành 10^{-5} và 10^{-3}. Tập dữ liệu nhỏ (S_1) được sử dụng cho quá trình huấn luyện và tập D_2^{Test} được sử dụng cho quá trình kiểm thử.

Các kết quả trình bày trong Bảng 3.9 cung cấp đánh giá định lượng về tác động của các tốc độ học khác nhau. Khi sử dụng tốc độ học cao *($Lr = 10e^{-3}$)*, mô hình thể hiện khả năng tổng quát hóa kém, phản ánh qua các chỉ số recall và AP thấp. Điều này có thể là do các cập nhật tham số quá lớn trong quá trình huấn luyện, dẫn đến sự mất ổn định hoặc không thể hội tụ đúng cách. Ngược lại, với tốc độ học thấp ($Lr = 10e^{-5}$), mặc dù mô hình duy trì được độ nhạy cao, nhưng sự sụt giảm nhẹ trong AP chỉ ra rằng tốc độ học có thể đang quá thấp, dẫn đến việc hội tụ chậm và tinh chỉnh tham số chưa đầy đủ.

Tốc độ học tối ưu được xác định là $Lr = 10e^{-4}$, tại đó mô hình đạt được sự cân bằng tốt nhất giữa việc cập nhật tham số và tính ổn định. Ở mức này, mô hình đạt được độ chính xác và độ nhạy cao trên tất cả các lớp tàu, khiến nó trở thành lựa chọn hiệu quả nhất trong kịch bản này.

Bảng 3.9: So sánh hiệu năng của mô hình Deformable DETR với learning rates khác nhau. Các kết quả tốt nhất được in đậm.

	$Lr = 10^{-3}$			$Lr = 10^{-4}$			$Lr = 10^{-5}$		
	Dets	**recall**	**AP**	**Dets**	**recall**	**AP**	**Dets**	**recall**	**AP**
fishing boat	75600	0,122	0,001	24874	**0,971**	**0,912**	15869	0,986	0,877
container ship	12600	0,223	0,007	9193	**0,993**	0,986	17028	0,988	0,968
ore carrier	12600	0,254	0,005	33544	**0,994**	0,965	24892	0,988	0,919
bulk cargo carrier	12600	0,401	0,016	35317	**0,995**	**0,949**	21334	0,997	0,906
passenger ship	14000	0,211	0,001	14543	**0,952**	0,861	42789	0,982	0,760
general cargo ship	12600	0,361	0,009	22529	**0,960**	**0,969**	18088	0,994	0,916
mAP			0,006			**0,941**			0,891

3.6.2 So sánh với các phương pháp tiên tiến nhất (SoTA)

Mục này so sánh phương pháp đề xuất với các phương pháp tiên tiến nhất (SoTA) dựa trên chỉ số mAP. Đối với mỗi phương pháp sẽ sử dụng tập dữ liệu huấn luyện và kiểm thử tương ứng. Cụ thể, chúng tôi sử dụng D^1_{Train} và D^1_{Test} để huấn luyện mô hình của mình và so sánh với Zhang_2022, Zhang_2021 [10], Liu_2020 , Liu_2022, Han_2021, SDNet_2022, và phương pháp dựa trên DETR. Ngoài ra, phương pháp đề xuất sử dụng D^2_{Train} và D^2_{Test} để huấn luyện một mô hình khác và so sánh với Biaohua_2022, Yani_2022 và phương pháp dựa trên DETR.

Bởi vì chỉ số mAP đạt cực đại khi $n_{queries} = 200$ đối với phương pháp của chúng tôi được thể hiện trong Bảng 3.8 , chúng tôi lựa chọn thiết lập này cho thí nghiệm này. Kết quả so sánh giữa các phương pháp SoTA được trình bày trong Bảng3.10 và một số kết luận có thể được rút ra như sau:

- Khung nền tảng (Baseline framework) là yếu tố then chốt để đạt được kết quả tốt hơn: Cui_2019 và Liu_2020 được phát triển dựa trên YOLOV3. Do đó, hiệu năng của chúng không tốt bằng Liu_2022, phương pháp vốn dựa trên khung SSD. Han_2021 dựa trên YOLOV4, nhưng hiệu năng không cho thấy sự cải thiện so với Liu_2020 (dựa trên YOLOV3). Tận dụng ưu thế của YOLOV5, SDNet_2022 đã cải thiện đáng kể so với Liu_2020. Khung YOLOV5 giúp chỉ số mAP tăng tới 8% so với khung YOLOV3. Trong khi một số phương pháp tiên tiến hiện nay sử dụng YOLOX làm nền tảng, phương pháp do chúng tôi đề xuất lại được xây dựng dựa trên mạng xương sống (backbone) DETR. Gần đây, các khung này đã trở thành những phương pháp ưu việt cho bài toán phát hiện đối tượng. Vì vậy, kết quả đạt được tốt hơn so với các phương pháp khác.Đáng chú ý là các nghiên cứu phát hiện tàu thường tận dụng một khung phát hiện đối tượng làm nền tảng, đi kèm với một số sửa đổi tùy chỉnh. Do đó, việc thừa hưởng các khả năng của một khung mới và mạnh mẽ như vậy tất yếu dẫn đến kết quả được cải thiện.
- Khi số lượng mẫu trong tập huấn luyện giảm, các phương pháp dựa trên DETR có xu hướng hoạt động hiệu quả hơn so với các phương pháp dựa trên CNN. Cụ thể,Trong bảng số liệu, Biaohua_2022 là mô hình phát hiện dựa trên CNN, còn Yani_2022 là mô hình dựa trên DETR. Chỉ số mAP của Yani_2022 và Biaohua_2022 lần lượt là 0,965 và 0,963. Tuy nhiên, hiệu năng này có thể được nâng cao nếu chúng ta lựa chọn siêu tham số phù hợp. Trong nghiên cứu của chúng tôi, bằng cách thiết lập $n_{queries} = 200$, chỉ số mAP đã được cải thiện lên mức 0,981.
- Phương pháp của chúng tôi tương đương với phương pháp dựa trên CNN tốt nhất hiện nay trong việc phát hiện tàu. Nếu tập dữ liệu huấn luyện có nhiều mẫu hơn tập kiểm thử, phương pháp của

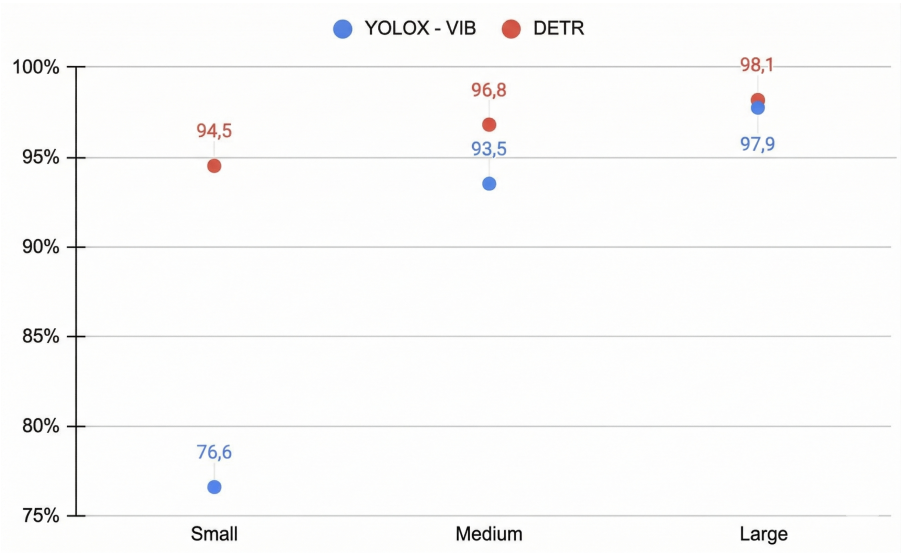
chúng tôi tương đương với phương pháp dựa trên YOLO như phương pháp dựa trên DETR.Cụ thể, khi sử dụng D^1_{Train} và D^1_{Test} cho quá trình huấn luyện và kiểm thử, chỉ số mAP của cả phương pháp của chúng tôi và phương pháp dựa trên DETR là tương tự nhau. Tuy nhiên, khi số lượng mẫu trong tập huấn luyện bằng với số lượng trong tập kiểm thử, phương pháp của chúng tôi tốt hơn một chút so với phương pháp dựa trên DETR

Bảng 3.10: So sánh hiệu suất của các phương pháp khác nhau. Kết quả tốt nhất được in đậm.

Method	**Train+Val / Test (in %)**	**fishing boat**	**container ship**	**ore carrier**	**bulk cargo carrier**	**passenger ship**	**general cargo ship**	**mAP**
Zhang_2022 [13]	90/10	0,824	0,940	0,859	0,915	0,787	0,914	0,873
Zhang_2021 [10]	90/10	-	-	-	-	-	-	0,946
Cui_2019 [9]	80/20	0,900	0,940	0,90	0,910	0,910	0,900	0,910
Liu_2020 [12]	80/20	-	-	-	-	-	-	0,908
Han_2021 [28]	80/20	-	-	-	-	-	-	0,906
Liu_2022 [7]	80/20	-	-	-	-	-	-	0,964
SDNet_2022 [11]	80/20	**0,986**	0,995	**0,989**	0,990	0,982	0,989	0,988
phương pháp dựa trên DETR [14]	80/20	0,979	**1**	0,987	**0,994**	0,994	**0,993**	**0,991**
Ours ($n_{query} = 200$)	80/20	0,982	**1**	**0,989**	0,991	**0,995**	0,990	**0,991**
Yani_2022 (ESDT) [17]	50/50	-	-	-	-	-	-	0,593
Yani_2022 (DETR) [17]	50/50	-	-	-	-	-	-	0,965
Biaohua_2022 [19]	50/50	0,940	0,987	0,966	0,978	0,937	0,972	0,963
phương pháp dựa trên DETR [14]	50/50	0,970	0,986	0,984	**0,991**	**0,964**	0,989	0,98
Ours ($n_{query} = 200$)	50/50	0,970	**0,995**	**0,992**	0,992	0,957	**0,992**	**0,982**

Trong Bảng 3.10, các kết quả chỉ ra rằng phương pháp DETR nhìn chung vượt trội hơn phương pháp dựa trên YOLOX khi số lượng mẫu huấn luyện bị giảm xuống. Do đó, chúng tôi đã thiết kế một thí nghiệm sử dụng D^2_{Test} làm tập kiểm thử và một tập con của D^2_{Train} làm tập huấn luyện. Các tập con này được phân loại thành các mức Nhỏ, Vừa và Lớn, tương ứng với các tập con S1, S2 và S3. Các kết quả trong Hình 3.18 so sánh phương pháp của chúng tôi với phương pháp tiên tiến nhất (SOTA) được đề xuất trong phương pháp dựa trên DETR. Nếu tập dữ liệu huấn luyện là S1, chỉ số mAP của phương pháp chúng tôi cải thiện 17,5% so với phương pháp dựa trên DETR. Khi số lượng mẫu huấn luyện tăng lên, mức độ cải thiện giảm đi. Mức tăng mAP là 3,3% nếu tập con S2 được sử dụng để huấn luyện; và nếu tập dữ liệu huấn luyện là S3, chỉ số mAP từ cả hai thiết lập là khá tương đương. Quan sát này chỉ ra rõ ràng rằng phương pháp dựa trên DETR có thể hữu ích nếu số lượng mẫu huấn luyện bị hạn chế. Bảng 3.10 báo cáo kết quả phát hiện chi tiết cho từng loại tàu.

Mặc dù các bộ phát hiện dựa trên DETR có thể đạt độ chính xác cao hơn trên các tập dữ liệu nhỏ, nhưng điều quan trọng là cũng cần đánh giá độ phức tạp tính toán của chúng so với các bộ phát hiện dựa trên CNN. Bảng 3.11 trình bày sự so sánh về hiệu năng tính toán giữa hai mô hình phát hiện đối



Hình 3.17: Hiệu suất khi số lượng mẫu huấn luyện bị giới hạn. "Small" nghĩa là 30% mẫu huấn luyện, "Medium" nghĩa là 70% mẫu huấn luyện, và "Large" nghĩa là 100% mẫu huấn luyện từ D_2^{Train}.

tượng: phương pháp dựa trên DETR và Deformable DETR. Ba chỉ số chính được so sánh bao gồm: Số khung hình trên giây, GFlops và số lượng tham số (tính bằng triệu).

Cả hai mô hình đều cho thấy chỉ số FPS tương đương nhau, trong đó DETR nhỉnh hơn phương pháp dựa trên DETR một chút (12,6 FPS so với 12,39 FPS trên GPU TiTanRTX). Tuy nhiên, DETR thể hiện độ phức tạp tính toán cao hơn, đòi hỏi 11,01 GFlops so với mức hiệu quả hơn là 9,92 GFlops của phương pháp dựa trên DETR. Mặc dù yêu cầu tính toán cao hơn, DETR lại sử dụng ít tham số hơn, với 39,82 triệu tham số so với 55,33 triệu của phương pháp dựa trên DETR. Điều này cho thấy DETR duy trì được tốc độ cạnh tranh với ít tham số hơn, nhưng phải đánh đổi bằng việc gia tăng độ phức tạp tính toán.

Trong khi các bộ phát hiện dựa trên DETR đạt độ chính xác cao trên tập dữ liệu nhỏ, chúng tôi cũng đánh giá độ phức tạp tính toán so với các phương pháp CNN. Bảng 3.12 trình bày so sánh giữa phương pháp dựa trên DETR và Deformable DETR.

Kết quả cho thấy cả hai mô hình có tốc độ xử lý (FPS) tương đương (12.6 so với 12,39 trên GPU TitanRTX). Mặc dù DETR có độ phức tạp tính toán cao hơn (11,01 GFlops so với 9,92 GFlops), nhưng lại sử dụng ít tham số hơn đáng kể (39,82 triệu so với 55,33 triệu). Điều này cho thấy DETR duy trì được tốc độ cạnh tranh với ít tham số hơn, dù phải đánh đổi bằng việc tăng nhẹ khối lượng tính toán.

Bảng 3.11: Hiệu suất trên các tập dữ liệu nhỏ. S_1 nghĩa là 30% mẫu huấn luyện, S_2 nghĩa là 70% mẫu huấn luyện, S_3 nghĩa là 100% mẫu huấn luyện từ D_2^{Train}.

Scenario	Metrics	fishing boat	container ship	ore carrier	bulk cargo carrier	passenger ship	general cargo ship	mAP
S_1 by VIB	recall	0,878	0,941	0,924	0,913	0,657	0,946	
	AP	0,796	0,884	0,831	0,765	0,524	0,794	0,766
S_1 by DETR	recall	0,971	0,993	0,994	0,995	0,952	0,96	
	AP	0,912	0,986	0,965	0,949	0,861	0,969	0,941
S_2 by VIB	recall	0,940	0,984	0,962	0,971	0,891	0,964	
	AP	0,922	0,980	0,935	0,953	0,873	0,946	0,935
S_2 by DETR	recall	0,977	1	0,992	0,995	0,968	0,997	
	AP	0,958	0,995	0,967	0,978	0,922	0,986	0,968
S_3 by VIB	recall	0,978	0,986	0,990	0,995	0,972	0,993	
	AP	0,970	0,986	0,984	0,991	0,964	0,989	0,98
S_3 by DETR	recall	0,985	0,995	0,999	0,998	0,980	0,997	
	AP	0,970	0,995	0,992	0,992	0,957	0,992	0,982

Bảng 3.12: So sánh độ phức tạp giữa VIB-detector và DETR-detector.

Chỉ số	phương pháp dựa trên DETR [14]	DETR (Ours)
FPS (Frames Per Second)	12,39	12,6
GFlops	9,92	11,01
#Parameters (Triệu tham số)	55,33	39,82

3.6.3 Nghiên cứu cắt giảm các hàm mất mát

Phần này thảo luận về một nghiên cứu cắt giảm đối với các hàm mất mát trong quá trình huấn luyện. Deformable DETR sử dụng nhiều hàm mất mát để huấn luyện, bao gồm focal loss, GIoU loss và L1 loss, mỗi hàm phục vụ một mục đích riêng biệt: focal loss dùng cho phân loại, GIoU dùng cho hồi quy hộp giới hạn và L1 dùng cho phát hiện đối tượng. Sự kết hợp của các hàm mất mát này đảm bảo sự thành công cho quá trình huấn luyện. Mặc dù tất cả các hàm mất mát đều đóng vai trò then chốt, nhưng mức độ đóng góp của chúng có thể được điều chỉnh. Theo mặc định, các trọng số được thiết lập ở mức 2.0 cho focal loss, 2.0 cho GIoU loss và 5.0 cho L1 loss. Để đánh giá tác động của từng thành phần, chúng tôi đã giảm trọng số của các hàm mất mát này đi 10 lần (thực hiện lần lượt từng hàm một) và so sánh kết quả với thiết lập mặc định. Các kết quả trong Bảng 3.13 chứng minh rằng L_{cls} (hàm mất mát phân loại) là yếu tố quan trọng nhất; việc giảm trọng số của nó dẫn đến sự sụt giảm hiệu năng đáng kể. Ngược lại, việc giảm hàm mất mát đối tượng gây ra ít tác động hơn, khi hiệu năng

vẫn duy trì ở mức gần tương đương với thiết lập ban đầu. Khả năng định vị bị ảnh hưởng nhẹ bởi hàm mất mát GIoU, thể hiện qua việc chỉ số mAP giảm xuống còn 0,931 so với 0,941 trong thiết lập ban đầu.

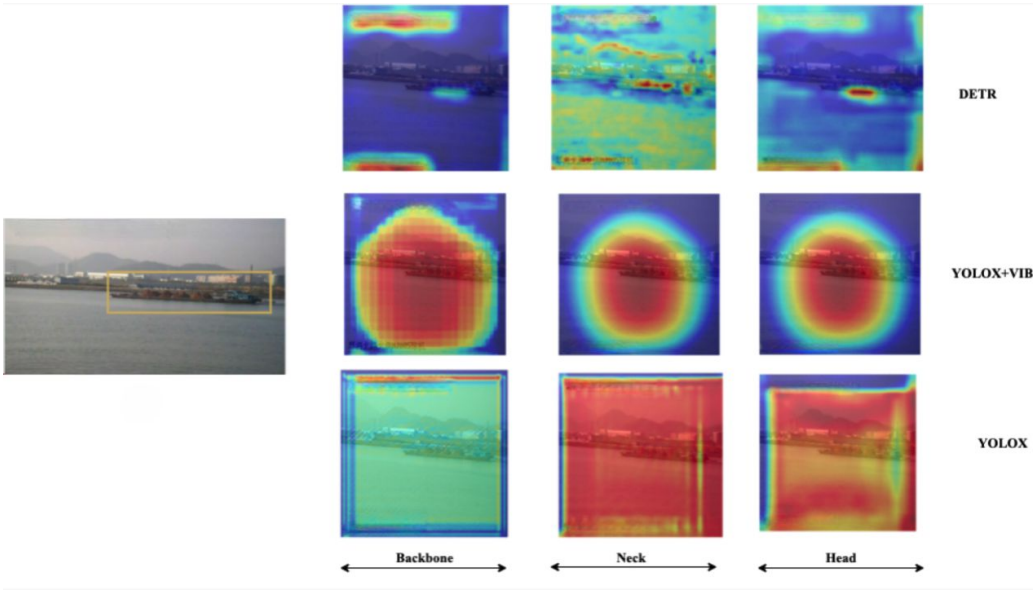
Bảng 3.13: So sánh mAP của Deformable DETR khi giảm các hàm mất mát huấn luyện.

	L_{GIoU}	L_1	L_{cls}
fishing boat	0,899	0,899	0,227
container ship	0,976	0,985	0,246
ore carrier	0,953	0,947	0,212
bulk cargo carrier	0,951	0,966	0,151
passenger ship	0,845	0,849	0,0173
general cargo ship	0,964	0,969	0,160
mAP	0,931	0,936	0,178

3.6.4 Phân tích đặc trưng

Các kết quả thực nghiệm trong Mục 3.6.2 chỉ ra rằng phương pháp DETR hoạt động tốt hơn các phương pháp CNN nếu số lượng mẫu trong tập huấn luyện bị hạn chế. Tuy nhiên, việc giải thích tại sao phương pháp DETR có thể đạt kết quả tốt hơn trong kịch bản đặc thù này là rất cần thiết. Sự khác biệt chính giữa hai phương pháp nằm ở mô-đun chú ý trong bộ mã hóa của DETR. Mô-đun này cho phép tương tác phi cục bộ để học các đặc trưng tốt hơn.

Do đó, chúng tôi trực quan hóa các đặc trưng được tạo ra bởi cả hai phương pháp sau các phần mạng cơ sở (backbone), lớp trung gian và tầng đầu ra (head) của mô hình. Với mỗi ảnh đầu vào, các bản đồ đặc trưng được trích xuất sau mỗi mô-đun. Tổng giá trị của một bản đồ đặc trưng đại diện cho mức độ quan trọng của bản đồ đó. Vì vậy, chúng tôi đã chọn 20 bản đồ đặc trưng quan trọng nhất cho mỗi phần mạng cơ sở, lớp trung gian và tầng đầu ra để tạo ra một bản đồ nhiệt. Bản đồ này được tính trung bình từ tất cả các bản đồ đặc trưng quan trọng và đại diện cho các điểm trọng yếu trên ảnh. Hình 3.18 minh họa các ví dụ về bản đồ nhiệt được tạo ra từ một ảnh đầu vào. Hàng đầu tiên hiển thị các bản đồ đặc trưng từ DETR, hàng thứ hai hiển thị các bản đồ nhiệt từ phương pháp phương pháp dựa trên DETR (phương pháp sử dụng hàm mất mát lựa chọn đặc trưng để học các đặc trưng), và hàng thứ ba trình bày các bản đồ nhiệt từ YOLOX vốn dựa hoàn toàn vào các mạng CNN. Các kết quả chỉ ra rằng DETR, với cơ chế chú ý của nó, có thể tập trung tốt hơn vào các đối tượng không phải là nền. Ví dụ, sau phần tầnng đầu ra*, bản đồ đặc trưng làm nổi bật phần văn bản hiển thị điểm số trên hệ thống sinh



Hình 3.18: Bản đồ đặc trưng tại đầu phân loại, phần cổ và phần xương sống. (Văn bản trên ảnh gốc đã được loại bỏ.)

tồn, và con tàu cũng được làm nổi bật, mặc dù không rõ ràng bằng phần văn bản. Khi các đặc trưng được xử lý qua lớp trung gian, các đặc trưng ngữ nghĩa cấp cao hơn được học, khiến con tàu trở nên nổi bật hơn trong khi sự tập trung vào văn bản giảm dần. Các điểm chính tập trung vào con tàu tại phần tầng đầu ra, với sự chú ý vào văn bản giảm đi.

Ngược lại, phương pháp dựa trên DETR tạo ra các bản đồ nhiệt thưa thớt, nơi nhiều điểm ảnh ở rìa bức ảnh không có phản hồi. Tuy nhiên, các bản đồ này không tập trung chính xác vào đối tượng. Điều này xảy ra bởi vì phương pháp dựa trên DETR sử dụng hàm mất mát chọn lọc đặc trưng để xác định các đặc trưng quan trọng, dẫn đến các bản đồ đặc trưng thưa thớt và có tính phân biệt cao nhưng không nằm chính xác ở tâm đối tượng. Sự phân bố bản đồ nhìn chung khá tương đồng, với những cải thiện nhỏ từ mạng cơ sở đến tầng đầu ra. Nếu không có hàm mất mát chọn lọc đặc trưng, sự phân bố của các bản đồ nhiệt có thể sẽ đồng đều hơn như được hiển thị ở hàng thứ ba.

3.7 Kết luận của chương

Trong chương này, luận án đã trình bày một cách toàn diện các kết quả thực nghiệm nhằm đánh giá hiệu quả của phương pháp phát hiện tàu biển được đề xuất dựa trên hai hướng tiếp cận chính: mô hình phát hiện dựa trên CNN (YOLOX kết hợp VIB) và mô hình phát hiện dựa trên transformer (DETR/Deformable DETR). Thông qua các thí nghiệm có kiểm soát chặt chẽ, nhiều khía cạnh quan trọng của mô hình đã được phân tích, bao gồm lựa chọn siêu tham số, ảnh hưởng của kiến trúc mạng,

vị trí tích hợp mô-đun, khả năng học với dữ liệu hạn chế, cũng như so sánh với các phương pháp tiên tiến nhất hiện nay.

Đối với mô hình YOLOX kết hợp với hàm mục tiêu VIB, kết quả thực nghiệm cho thấy việc lựa chọn siêu tham số điều khiển thành phần KL đóng vai trò then chốt trong việc cân bằng giữa khả năng học đặc trưng và nhiệm vụ phát hiện chính. Khi giá trị của tham số này quá nhỏ, mô hình không tận dụng được lợi ích của việc lựa chọn đặc trưng; ngược lại, khi giá trị quá lớn, quá trình rút trích đặc trưng trở nên quá khắt khe, dẫn đến việc loại bỏ thông tin cần thiết cho phân loại và làm suy giảm hiệu suất. Giá trị tối ưu được xác định thông qua thực nghiệm cho phép mô hình đạt độ chính xác cao nhất, đồng thời đảm bảo tính ổn định trong huấn luyện.

Các thí nghiệm với nhiều mạng xương sống khác nhau cho thấy phương pháp đề xuất có tính tổng quát tốt và không phụ thuộc chặt chẽ vào một kiến trúc trích xuất đặc trưng cụ thể. Trong đó, DarkNet tiếp tục thể hiện ưu thế khi được sử dụng trong khung YOLO, trong khi ResNet và MobileNet cũng thu được mức cải thiện đáng kể khi tích hợp hàm mục tiêu VIB. Điều này khẳng định rằng cơ chế lựa chọn đặc trưng dựa trên VIB có khả năng hỗ trợ các mạng xương sống khác nhau bằng cách làm nổi bật thông tin liên quan đến đối tượng và giảm ảnh hưởng của nền.

Phân tích trực quan các bản đồ đặc trưng cung cấp bằng chứng định tính rõ ràng cho hiệu quả của phương pháp. Khi sử dụng VIB, các phản hồi đặc trưng tập trung mạnh vào vùng chứa tàu, trong khi các mô hình không sử dụng VIB có xu hướng phân tán phản hồi trên toàn ảnh. Hiện tượng này không chỉ xuất hiện ở nhánh phân loại mà còn lan truyền ngược đến các khối kết hợp đặc trưng, cho thấy tác động tích cực của hàm mục tiêu VIB đối với toàn bộ quá trình học biểu diễn.

Về vị trí tích hợp mô-đun VIB, các thí nghiệm cho thấy việc chèn VIB vào nhánh phân loại là lựa chọn phù hợp nhất. Khi mô-đun này được đưa vào nhánh hồi quy hoặc đồng thời vào cả hai nhánh, quá trình huấn luyện trở nên không ổn định và mô hình không thể hội tụ. Phân tích các hàm mất mát trong quá trình huấn luyện chỉ ra rằng việc áp đặt ràng buộc KL lên không gian đặc trưng của nhánh hồi quy làm gia tăng độ khó của bài toán dự đoán hộp bao, từ đó ảnh hưởng tiêu cực đến toàn bộ hệ thống. Kết quả này mang ý nghĩa thực tiễn quan trọng, cung cấp hướng dẫn rõ ràng cho việc tích hợp các mô-đun lựa chọn đặc trưng vào các mô hình phát hiện đối tượng trong tương lai.

So sánh với các phương pháp tiên tiến nhất trong phát hiện tàu biển cho thấy phương pháp đề xuất đạt hiệu suất cạnh tranh và trong nhiều trường hợp vượt trội, đặc biệt khi số lượng mẫu huấn luyện bị hạn chế. Việc tái lấy mẫu trong không gian đặc trưng thông qua VIB giúp mô hình tăng khả năng khái

quát hóa và giảm hiện tượng quá khớp, nhất là trong bối cảnh dữ liệu huấn luyện nhỏ. Điều này được thể hiện rõ ràng qua các thí nghiệm với các tập con dữ liệu có quy mô khác nhau, trong đó mức cải thiện mAP lớn nhất đạt được trên tập dữ liệu nhỏ nhất và giảm dần khi số lượng mẫu huấn luyện tăng lên.

Đối với các mô hình dựa trên DETR, chương này đã phân tích chi tiết ảnh hưởng của các siêu tham số quan trọng như số lượng truy vấn và tốc độ học. Kết quả cho thấy việc lựa chọn số lượng truy vấn phù hợp không chỉ giúp cải thiện độ chính xác mà còn giảm sự mất cân bằng trong số lượng phát hiện giữa các lớp tàu. Đồng thời, tốc độ học tối ưu giúp mô hình đạt được sự cân bằng giữa khả năng hội tụ và tính ổn định. Khi được cấu hình đúng, mô hình dựa trên DETR thể hiện ưu thế rõ rệt trong các kịch bản dữ liệu hạn chế, vượt trội hơn các phương pháp dựa trên CNN truyền thống.

Các thí nghiệm so sánh trực tiếp giữa phương pháp đề xuất và các phương pháp SoTA cho thấy rằng, khi dữ liệu huấn luyện dồi dào, hiệu suất của các mô hình dựa trên CNN và transformer là tương đương. Tuy nhiên, khi số lượng mẫu huấn luyện giảm, các mô hình dựa trên transformer, đặc biệt là phương pháp đề xuất, cho thấy khả năng thích nghi tốt hơn và duy trì độ chính xác cao hơn. Phân tích độ phức tạp tính toán cũng chỉ ra rằng mặc dù các mô hình transformer có chi phí tính toán cao hơn, chúng vẫn đạt được tốc độ xử lý cạnh tranh và sử dụng số lượng tham số hợp lý, qua đó mở ra tiềm năng ứng dụng trong các hệ thống giám sát biển thực tế.

Bên cạnh những kết quả tốt, việc phân tích các lỗi sai cũng cho thấy mô hình còn một số hạn chế khi gặp tình huống phức tạp. mô hình chủ yếu nhận diện nhầm khi các vật thể trên nền ảnh hoặc ven bờ (như mỏm đá nhỏ, phao, hoặc công trình ở cảng) có hình dáng giống với tàu thuyền. Ngược lại, hệ thống hay bỏ sót các tàu quá nhỏ ở xa, hoặc những tàu bị che khuất nhiều. Khi đánh giá hiệu quả trên từng loại tàu, độ chính xác ở một số loại (ví dụ như tàu chuyên dụng) bị giảm xuống. Nguyên nhân chính là do số lượng ảnh để huấn luyện không đồng đều và hình dáng bên ngoài của các loại tàu này lại khá giống nhau. Dù đã dùng các kỹ thuật cải tiến để máy học cách phân biệt tốt hơn, nhưng việc phân loại chi tiết khi thiếu ảnh mẫu vẫn là một bài toán khó cần khắc phục. Thêm vào đó, việc thử nghiệm trên nhiều điều kiện môi trường khác nhau cũng cho thấy rõ độ ổn định và cả giới hạn của hệ thống. Vào ban ngày và thời tiết đẹp, mô hình khoanh vùng tàu rất chính xác. Tuy nhiên, dưới thời tiết xấu như sương mù, mưa lớn hoặc vào ban đêm, hình ảnh bị mờ và kém tương phản. Điều này làm mất đi các đường nét rõ ràng, khiến mô hình dự đoán kém tự tin đi và dễ bỏ sót tàu hơn Tổng hợp lại, các kết quả thực nghiệm trong chương này đã chứng minh tính hiệu quả, ổn định và khả năng tổng quát của phương pháp phát hiện tàu biển được đề xuất. Những phân tích cả định lượng lẫn định tính không

chỉ xác nhận các đóng góp về mặt thuật toán mà còn cung cấp những hiểu biết sâu sắc về cách thức thiết kế, huấn luyện và triển khai các mô hình phát hiện đối tượng trong điều kiện dữ liệu hạn chế. Đây là nền tảng quan trọng cho các nghiên cứu tiếp theo cũng như cho việc ứng dụng thực tế trong các hệ thống giám sát và an ninh hàng hải.

Chương 4

PHÂN TÍCH DỮ LIỆU RADAR BẰNG KỸ THUẬT HỌC SÂU

Nhận dạng và phân loại tàu biển dựa trên dữ liệu ảnh có ưu điểm trực quan, nhận biết rõ được kiểu loại, hành động của tàu nhưng camera bị hạn chế bởi cự ly quan sát không được xa và chịu ảnh hưởng nhiều bởi yếu tố thời tiết như mưa, sương mù... Radar có ưu điểm quan sát được ở cự ly xa,ít phụ thuộc vào thời tiết, có thể quan sát 24h. Chính vì vậy tại các trạm radar thường kết hợp giúp quan sát bằng camera và radar. Chương 4 nghiên cứu bài toán nhận dạng và phân loại tàu biển dựa trên tín hiệu radar xung thu thập thực tế. Đặc điểm tín hiệu phản xạ của các nhóm tàu khác nhau được phân tích nhằm xây dựng tập đặc trưng phù hợp. Luận án đề xuất phương pháp kết hợp tri thức chuyên gia với mô hình học sâu để rút trích đặc trưng có ý nghĩa vật lý và tính phân biệt cao. Các kỹ thuật phân cụm và phân loại được triển khai nhằm đánh giá khả năng tách lớp mục tiêu. Kết quả thực nghiệm chứng minh hiệu quả của phương pháp đề xuất so với các phương pháp truyền thống, đồng thời khẳng định giá trị của dữ liệu radar thực địa trong bài toán cảnh giới tự động.

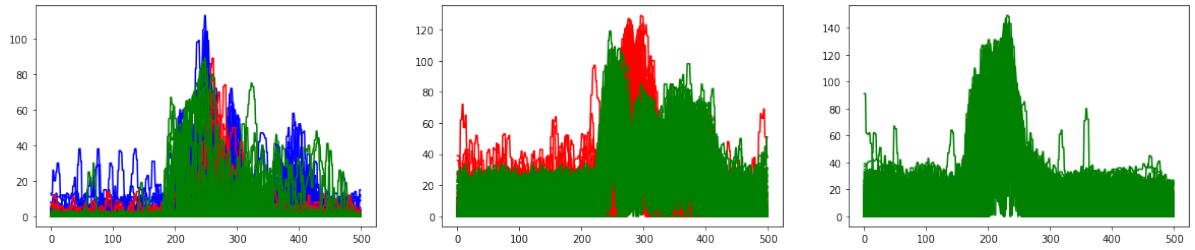
4.1 Phân tích đặc trưng tín hiệu Radar và phương pháp phân cụm

Radar đóng vai trò then chốt trong hệ thống giám sát an ninh hàng hải nhờ khả năng hoạt động liên tục trong mọi điều kiện thời tiết. Trong phạm vi nghiên cứu này, Nghiên cứu sinh tập trung khai thác dữ liệu từ radar xung. Đây là loại khí tài phổ biến trong quân sự và cảnh giới biển nhờ ưu điểm về công suất đỉnh lớn, khả năng quan sát tầm xa và tính bảo mật cao thông qua cơ chế phát xạ gián đoạn.

4.1.1 Thách thức trong phân tích dữ liệu radar thực tế

Dựa trên bộ dữ liệu radar xung thu thập tại vùng biển Việt Nam, Nghiên cứu sinh đã tiến hành khảo sát các đặc trưng phản xạ của ba loại mục tiêu chính: tàu cá, tàu vận tải và tàu quân sự. Các thử nghiệm phân cụm sơ bộ sử dụng thuật toán K-means cho thấy một thách thức lớn: dữ liệu có xu hướng bị gom nhóm dựa trên biên độ tín hiệu (phụ thuộc vào khoảng cách tới đài radar) thay vì dựa trên đặc trưng Hình thái của tàu.

Như minh họa tại Hình 4.1, các cụm dữ liệu thu được có sự chồng lấn đáng kể giữa các loại tàu. Tàu vận tải (kích thước lớn, nhiều container) và tàu cá (vỏ gỗ, kích thước nhỏ) thường bị phân loại sai nếu chỉ dựa vào cường độ phản xạ đỉnh, do biên độ tín hiệu bị suy hao mạnh theo cự ly. Thực tế này đặt ra yêu cầu cấp thiết phải trích xuất được các đặc trưng bất biến với khoảng cách, tập trung vào cấu trúc Hình dáng của xung phản hồi.



(a) Nhóm 1 (b) Nhóm 2 (c) Nhóm 3

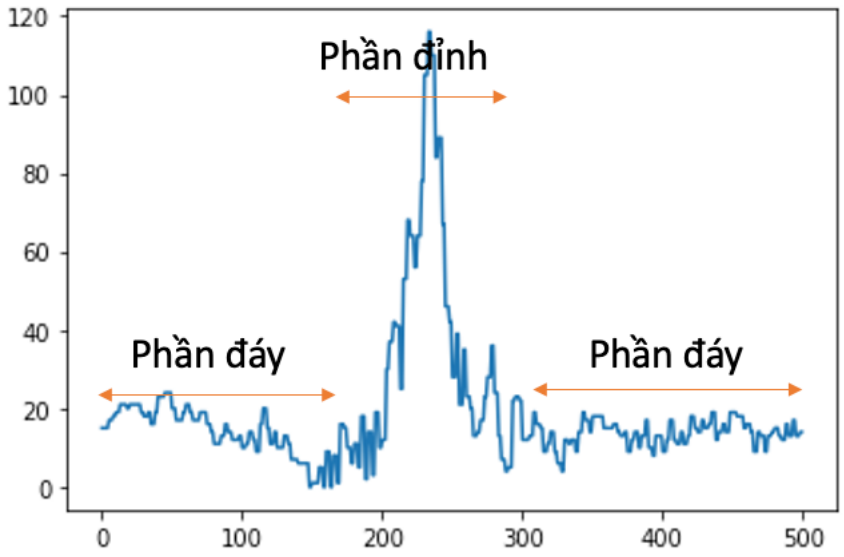
Hình 4.1: Ba nhóm dạng sóng thu được từ thực địa cho thấy sự chồng lấn về biên độ.

4.1.2 Đề xuất phương pháp trích xuất đặc trưng kết hợp

Để giải quyết hạn chế của việc tín hiệu bị phụ thuộc vào khoảng cách, luận án đề xuất phương pháp trích xuất đặc trưng lai, kết hợp chặt chẽ giữa khả năng học biểu diễn của mô hình học sâu Autoencoder và tri thức chuyên gia trong lĩnh vực xử lý tín hiệu radar.

Thay vì để mạng nơ-ron học đặc trưng một cách tự do theo dạng hộp đen, phương pháp đề xuất đưa vào các ràng buộc vật lý dựa trên kinh nghiệm thực tiễn và lý thuyết tán xạ điện từ của trắc thủ radar. Cụ thể, dạng sóng phản hồi được phân rã thành hai vùng Hình thái học cốt lõi là phần đỉnh sóng và phần đáy sóng (Hình 4.2). Tại các vùng này, tri thức chuyên gia được lượng hóa thành các thông số vật lý mang tính bất biến với khoảng cách. Đầu tiên là độ dốc sườn xung, được tính toán thông qua đạo hàm bậc nhất của tín hiệu tại sườn lên và sườn xuống. Đại lượng này phản ánh trực tiếp mức độ góc cạnh và đặc tính phản xạ bề mặt của vật liệu làm vỏ tàu. Tiếp theo là độ rộng xung tại các ngưỡng

năng lượng, đo lường độ trải dãn thời gian của xung tại các mức suy giảm cụ thể như -3dB và -10dB, cung cấp thông tin tỷ lệ thuận với kích thước vật lý dọc theo trục bức xạ của mục tiêu. Cuối cùng là tỷ số tín hiệu trên nhiễu cục bộ và phương sai nhiễu tại đáy sóng, giúp mô hình hóa sự nhiễu loạn bề mặt và các cấu trúc tán xạ phụ của thân tàu, vốn là những đặc trưng khó bị sai lệch bởi sự suy hao năng lượng trên đường truyền.



Hình 4.2: Phân tách đặc trưng phần đỉnh và phần đáy của xung phản hồi dựa trên tri thức chuyên gia.

mô hình Autoencoder [68] sau đó được thiết kế để nhận đầu vào là sự kết hợp của các vector thông số vật lý chuyên gia này cùng với chuỗi tín hiệu nguyên bản. Quá trình này giúp giảm chiều dữ liệu và mã hóa các đặc trưng lai vào một không gian tiềm ẩn cô đọng. Việc ép mô hình chú ý vào các quy tắc vật lý thiết yếu giúp các đặc trưng sau khi mã hóa trở thành đầu vào tối ưu cho thuật toán phân cụm K-means, từ đó tách biệt chính xác các nhóm tàu dựa trên bản chất kiến trúc và vật liệu thay vì sự thay đổi biên độ đơn thuần.

4.1.3 Mục tiêu nghiên cứu của chương

Dựa trên cách tiếp cận trên, chương này tập trung giải quyết ba nhiệm vụ chính:

1. Xây dựng và chuẩn hóa bộ dữ liệu radar xung từ thực địa tại Việt Nam.
2. Phân tích và đánh giá hiệu quả của các đặc trưng chuyên gia thông qua mô hình phân cụm học sâu (Deep Clustering).
3. Đề xuất và huấn luyện bộ phân loại tự động nhằm định danh chính xác loại tàu từ tín hiệu radar

phản hồi.

4.2 Phân đoạn và trích xuất đặc trưng

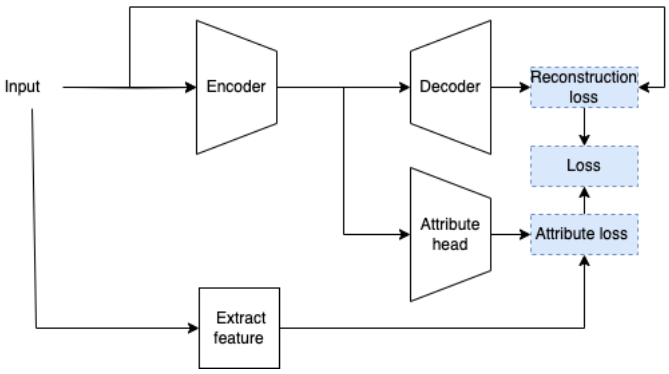
Mặc dù phương pháp phân loại tàu biển dựa trên học sâu sẽ được trình bày chi tiết ở phần sau, việc thực hiện phân cụm tín hiệu radar ở bước này đóng vai trò là cơ sở tiền đề mang tính quyết định. Do tín hiệu radar xung ở dạng thô thường chứa nhiều nhiễu và có tính trừu tượng cao, việc đưa trực tiếp vào mạng nơ-ron sẽ gây khó khăn cho quá trình hội tụ. Vì vậy, việc phân cụm dựa trên các thuộc tính do chuyên gia đề xuất giúp phân hoạch không gian đặc trưng, gom nhóm các mẫu tín hiệu có chung bản chất vật lý (như cấu trúc vỏ tàu, kích thước).

Quá trình này giúp làm giảm độ biến thiên của dữ liệu, tạo ra các vector đặc trưng có tính phân biệt cao, từ đó làm nền tảng vững chắc giúp mô hình phân loại phía sau dễ dàng phân tách các lớp, hội tụ nhanh hơn và nâng cao hiệu năng nhận dạng tổng thể.

4.2.1 Tổng quan về hệ thống

Quy trình phân đoạn tín hiệu radar tại Chương 4 được thực hiện qua chuỗi xử lý nối tiếp nhau. Đầu tiên, tín hiệu thô được tiền xử lý làm sạch để trích xuất các đặc trưng phổ radar. Hệ thống sau đó áp dụng chiến lược phân cụm kết hợp: sử dụng K-means để khởi tạo các tâm cụm ban đầu, làm nền tảng cho các mô hình học sâu (DEC, DeepCluster) học, tinh chỉnh không gian biểu diễn và phân tách chính xác các nhóm tín hiệu mục tiêu phức tạp. Chi tiết đường đi của dữ liệu và vai trò của từng thuật toán được hệ thống hóa trực quan trong sơ đồ luồng xử lý tổng thể đi kèm. Tổng quan vệ hệ thống được trình bày như trong Hình 4.3. Theo đó, bộ dữ liệu về tín hiệu radar được gán các thuộc tính theo kiến thức của chuyên gia. Sau đó mô hình học sâu sẽ được học các đặc trưng để vừa tái tạo lại tốt thông tin của sóng radar phản hồi, vừa dự đoán các thuộc tính được gán nhãn cho mỗi mẫu. Cuối cùng, các đặc trưng đó sẽ được sử dụng để phân cụm dữ liệu.

Phần 4.2.2, trình bày cách các thuộc tính mô tả các loại tàu biển và nguyên nhân lựa chọn các thuộc tính ấy. Phần 4.2.3 mô tả cách gán thuộc tính cho mỗi mẫu trong bộ dữ liệu. Phần 4.2.4 trình bày kiến trúc của model và hàm mục tiêu.



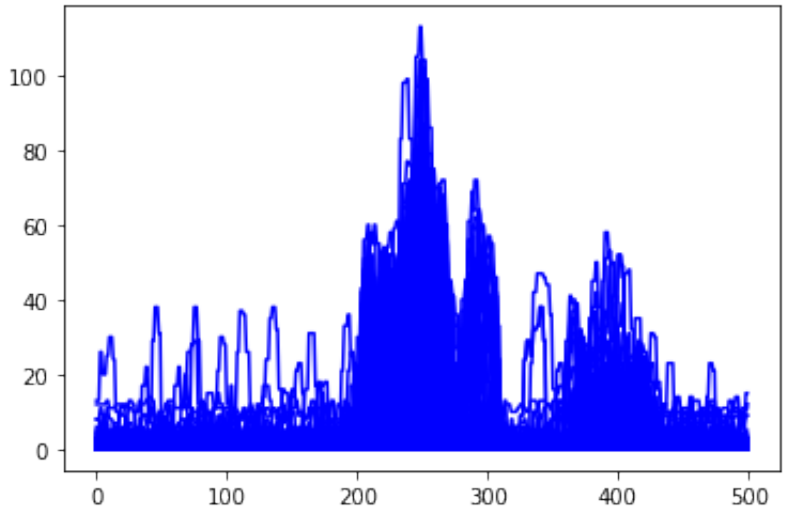
Hình 4.3: Tổng quan về các phương pháp được đề nghị.

4.2.2 Các thuộc tính được đề xuất bởi chuyên gia

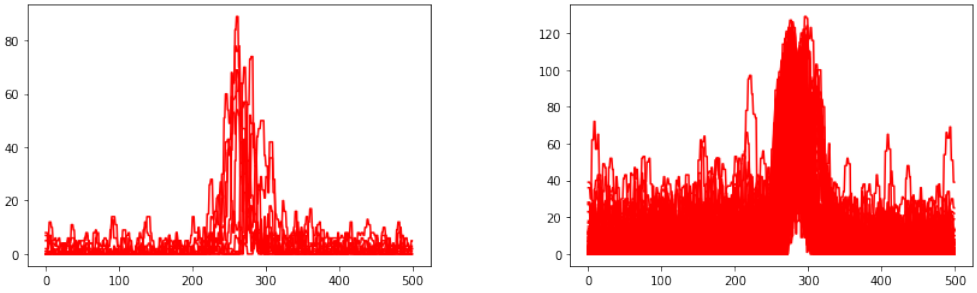
Để có thể phân tích các thuộc tính có thể hỗ trợ cho việc phân cụm tàu biển, các mẫu tàu cá, tàu vận tải và tàu quận sự đã được phân tách và hiển thị như trong Hình 4.4, 4.5, 4.6. Các tàu cá có kích thước nhỏ và đặc điểm vỏ tàu khiến sóng phản hồi có biên độ nhỏ. Bên cạnh đó do các tàu cá thường đi cùng với nhau nên có thể có nhiều đỉnh sóng trong một sóng phản hồi như trong Hình 4.4. Trong Hình 4.5, ta thấy nếu tàu quân sự ở xa trạm radar thì sóng phản hồi sẽ có biên độ đỉnh sóng nhỏ như trong Hình 4.5a; ngược lại nếu tàu quân sự ở gần trạm radar thì sóng phản hồi sẽ có biên độ đỉnh sóng và biên độ đáy sóng lớn như trong Hình 4.5b. Đối với tàu vận tải; nếu tàu vận tải đang ở xa và có xu hướng chạy ngang bờ biển, thì sóng phản hồi có biên độ thấp và độ rộng đỉnh sóng cao hơn do có ảnh hưởng của các thùng hàng như như Hình 4.6a. Nếu tàu vận tải chạy gần bờ nhưng không hướng vào cảng, thì sóng phản hồi có biên độ lớn hơn, và phần đáy sóng cũng có biên độ lớn; ngoài ra phần nhấp nhô của phần đỉnh sóng chuyển về vẫn lớn như trong Hình 4.6b. Nếu tàu vận tải chạy gần bờ nhưng có xu hướng hướng vào cảng, thì biên độ sóng lớn, nhưng độ rộng của sóng chỉ ở trung bình do lúc này chỉ có phần mũi tàu tiếp xúc với sóng radar như trong Hình 4.6c.

Từ những phân tích trên, các đặc điểm sau sẽ được sử dụng để đánh giá loại tàu dựa vào sóng phản hồi của radar.

- Số lượng đỉnh của sóng phản hồi: thông thường, dạng sóng phản hồi từ radar chỉ có một đỉnh nhưng đối với tàu cá, do đối tượng nhỏ và đôi khi bị nhiễu bởi các tàu cá lên cận nên có thể xuất hiện nhiều hơn một đỉnh. Nhưng nHình chung số lượng đỉnh không quá nhiều đối với một sóng phản hồi.
- Biên độ phần đỉnh của sóng phản hồi: Đây là giá trị lớn nhất của phần đỉnh sóng, thường ảnh hưởng rất lớn tới kết quả phân cụm nếu chỉ sử dụng tín hiệu cảm biến thô.



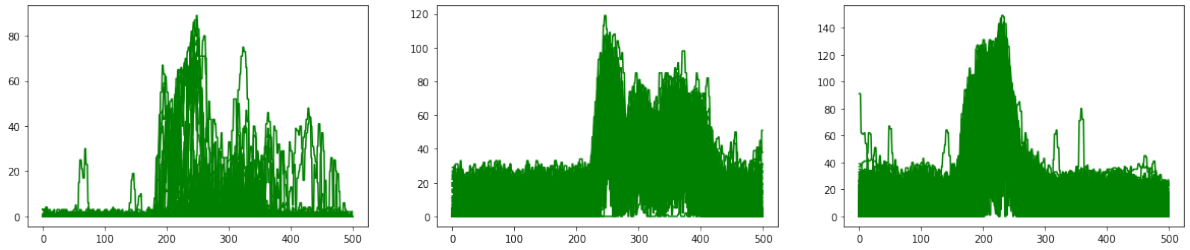
Hình 4.4: Dạng sóng phản hồi đặc trưng của tàu cá (biên độ nhỏ, thường xuất hiện nhiều đỉnh sóng do đi theo cụm).



(a) Tàu ở xa trạm radar (biên độ đỉnh sóng nhỏ).

(b) Tàu ở gần trạm radar (biên độ đỉnh và đáy sóng lớn).

Hình 4.5: Dạng sóng phản hồi của tàu quân sự theo các khoảng cách khác nhau.



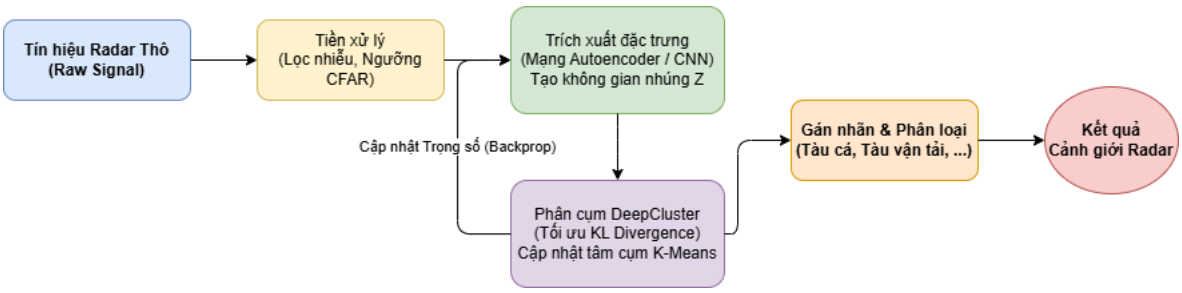
(a) Tàu ở xa, chạy ngang bờ (độ rộng đỉnh sóng cao)

(b) Tàu ở gần, chạy ngang bờ (biên độ và đáy sóng lớn)

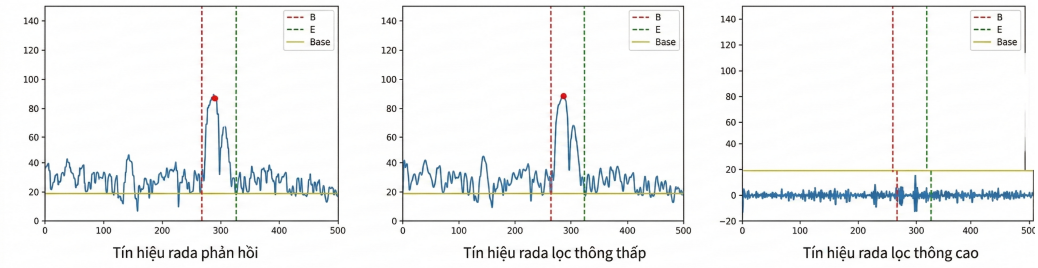
(c) Tàu ở gần, hướng vào cảng (độ rộng sóng trung bình)

Hình 4.6: Dạng sóng phản hồi của tàu vận tải theo hướng di chuyển và khoảng cách.

- Độ rộng phần đỉnh của sóng phản hồi: đặc trưng này rất nhỏ với tín hiệu phản hồi từ tàu cá. Đối với tàu quân sự, giá trị này có thể nhỏ hoặc trung bình tuỳ thuộc vào khoảng cách tàu. Trong khi đó, độ rộng đỉnh sóng của tàu vận tải rất lớn do đặc thù nhiều thùng hàng được xếp lên tàu. Mặc dù vậy, đặc trưng này chỉ nhận giá trị trung bình nếu tàu hướng vào bờ biển.



Hình 4.7: lưu đồ minh hoạ cho phương pháp xử lý tín hiêu radar.



Hình 4.8: Tiền xử lý tín hiệu radar để rút trích đặc trưng.

- Độ dao động phần đỉnh của sóng phản hồi: Đặc trưng này mô tả độ dao động ở phần đỉnh của dạng sóng. Nếu là tàu cá hoặc tàu quân sự, giá trị này có xu hướng nhỏ hơn. Trong khi tàu vận tải thường có giá trị lớn khi mô tả bằng đặc trưng này.
- Biên độ phần đáy của sóng phản hồi: Đặc trưng này mô tả phần đáy của một dáng sóng. Nếu là tàu cá, giá trị này rất nhỏ vì kích thước của tàu cá nhỏ. Tàu vận tải hoặc tàu quân sự có giá trị lớn hơn vì đặc thù của vỏ tàu.
- Độ dao động phần đáy của sóng phản hồi: Với đặc trưng này, tàu cá thường có giá trị nhỏ, và tàu quận sự, vận tải thường có giá trị lớn.

4.2.3 Rút trích đặc trưng

Gọi tín hiệu sóng radar phản hồi về là x, quy trình rút trích các thuộc tính trong phần 4.2.2, được mô tả như sau:

1. Áp dụng bộ lọc Butterworth bậc 1 để rút trích thành phần thông thấp x^H và thông cao x^L. Một ví dụ về ba tín hiêu x, x^H, và x^L được mô tả như trong Hình 4.8.
2. Áp dụng thuật toán phát hiện đỉnh [69] để phát hiện số đỉnh n_{peak} của dạng sóng x.
3. Sử dụng tín hiệu tần số thấp x^L, có thể tách phần đỉnh và đáy của dạng sóng dựa vào thời điểm

bắt đầu B và kết thúc E của đỉnh sóng cao nhất. Mô tả của các thời điểm B và E được thể hiện như trong Hình 4.8. Gọi $min_{30}(x)$ là tập hợp 30% số đỉnh có biên độ nhỏ nhất của tín hiệu x, giá trị đáy *Base* trong Hình 4.8 được xác định bằng công thức $Base = mean(min_{30}(x))$.

4. Dựa vào các thời điểm thời điểm bắt đầu B và kết thúc E của đỉnh sóng cao nhất, rút trích thông tin phần đỉnh và phần đáy của thành phần tần số thấp x^L và thành phần tần số cao x^H. Cụ thể x^L_{BE} là phần đỉnh của tín hiệu tần số thấp; trong khi $x^L_{\sim BE}$ là phần đáy của tính hiệu tần số thấp.

5. Gọi N là số phần tử trong một mẫu tín hiệu radar phản xạ. Các đặc trưng về độ rộng phần đỉnh sóng, biên độ phần đỉnh sóng, độ dao động phần đỉnh sóng, biên độ phần đáy sóng, độ dao động phần đáy sóng được tính bằng các phương trình 4.1, 4.2, 4.3, 4.4, và 4.5 một cách tương ứng.

$$L_{peak} = \frac{E-B}{N} \tag{4.1}$$

$$A_{peak} = mean(max_{30}(x^L_{BE})) \tag{4.2}$$

$$S_{peak} = mean[{x^H_{BE}}^2] \tag{4.3}$$

$$A_{base} = mean(x_{\sim BE}) \tag{4.4}$$

$$S_{base} = mean[{x^H_{\sim BE}}^2] \tag{4.5}$$

4.2.4 mô hình rút trích đặc trưng và hàm mục tiêu

Kiến trúc của mô hình sử dụng trong bài báo được mô tả như sau:

mô hình encoder: $f_{\theta^E}(,)$ mô hình này được sử dùng để rút trích đặc trưng. Khối này là sự nối tiếp của các khối trong danh sách $[nn, linear, nn, relu,]$.

mô hình decoder: $f_{\theta^D}(,)$ mô hình này được sử dùng để khôi phục lại thông tin gốc. Khối này là sự nối tiếp của các khối trong danh sách $[nn, linear, nn, relu,]$.

mô hình dự đoán thuộc tính $f_{\theta^A}(,)$**:** Khối này dùng để dự đoán các thuộc tính từ các đặc trưng được

rút trích từ khối encoder. Các lớp trong khối dự đoán thuộc tính được mô tả như sau $[nn, linear, nn, relu]$

Gọi x là vector 500 phần tử mô tả tín hiệu vào của cảm biến và $\hat{x}$ là dạng sóng được khôi phục lại, hàm mục tiêu khôi phục dữ liệu được mô tả như phương trình 4.6.

$$L_R(x,\hat{x}) = ||x - \hat{x}||^2 \tag{4.6}$$

Goi a là thuộc tính được gắn nhãn từ phần 4.2.3 và $\hat{a}$ là thuộc tính được dự đoán từ mô hình dự đoán thuộc tính. Hàm mục tiêu xác thuộc tính được mô tả như phương trình 4.7.

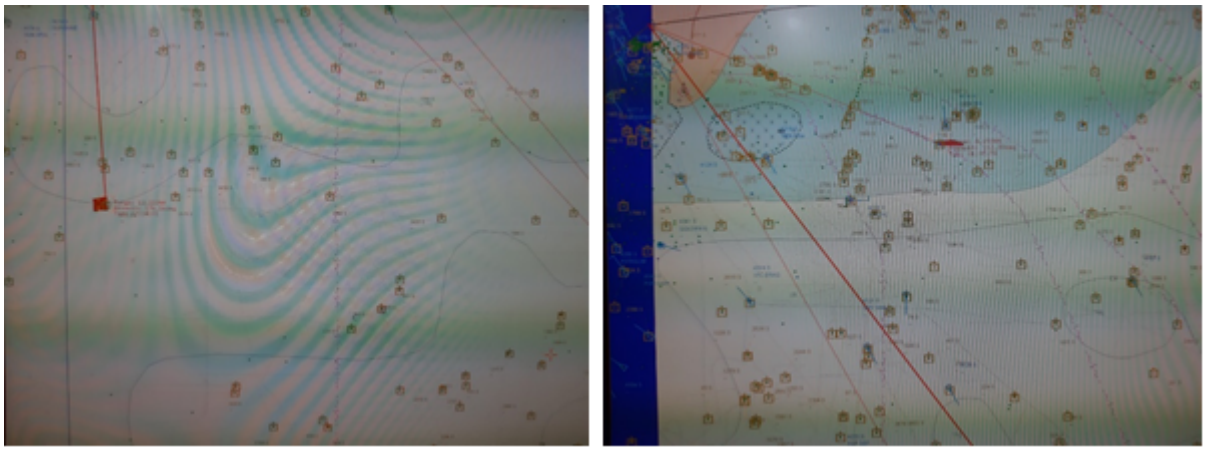
$$L_A(a,\hat{a}) = ||a - \hat{a}||^2 \tag{4.7}$$

Tham số α điều khiển sự cân bằng của hai ràng buộc $L_R(x,\hat{x})$ và $L_A(a,\hat{a})$. Kết quả hàm mục tiêu để huấn luyện bộ rút trích đặc trưng sẽ được mô tả như sau:

$$Loss(x,a) = L_R(x, f_{\theta^D}(f_{\theta^E}(x))) + \alpha L_A(a, f_{\theta^A}(f_{\theta^E}(x))) \tag{4.8}$$

4.2.5 Bộ dữ liệu và quá trình thu thập dữ liệu

Phần này của cuốn luận án tập trung giới thiệu quy trình thu nhận dữ liệu radar giám sát bờ biển. Thông thường, trạm radar cảnh giới bờ biển được bố trí và lắp đặt trải dài theo bờ biển Việt Nam. Hệ thống được lắp đặt sao cho mỗi trạm radar sẽ quan sát một phạm vi bờ biển nhất định và tất cả các trạm trong hệ thống sẽ quan sát tổng thể phạm vi bờ biển của Việt Nam.

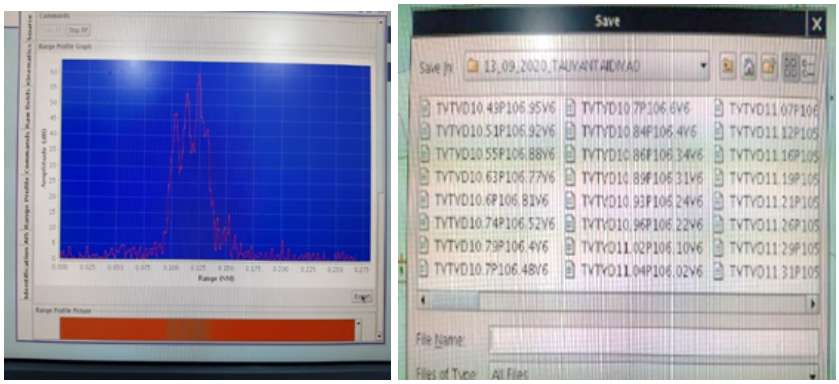


Hình 4.9: Màn Hình điều khiển khi xuất hiện mục tiêu

Quy trình thu thập và tiền xử lý dữ liệu Radar được thực hiện nghiêm ngặt nhằm đảm bảo tính trung thực và độ tin cậy của tín hiệu đầu vào. Tín hiệu radar phản hồi từ mục tiêu tàu biển trong môi trường cảnh giới thực tế ban đầu tồn tại dưới dạng sóng thô với độ phức tạp cao và chứa nhiều thành phần nhiễu mặt biển. Để trích xuất được thông tin hữu ích, các tín hiệu này được đưa vào hệ thống phần mềm xử lý chuyên dụng nhằm thực hiện các thuật toán lọc nhiễu, chuẩn hóa biên độ và tách biệt thành phần Hình thái học. Tín hiệu sau khi được làm sạch và tối ưu hóa sẽ được trích xuất sang định dạng dữ liệu cấu trúc (.txt). Các tệp tin này lưu trữ các vector giá trị số mô tả chi tiết đặc tính biên độ và pha của sóng radar theo thời gian, đóng vai trò là nguồn dữ liệu chuẩn hóa để huấn luyện và kiểm chứng các mô hình học máy trong các nội dung tiếp theo của luận án.

Do các hệ thống thu thập dữ liệu radar không hỗ trợ để thu dữ liệu tự động. Chính vì vậy Nghiên cứu sinh phải tiến hành thu dữ liệu thủ công. Việc thu thập dữ liệu tại một trạm radar được tiến hành như sau. Đầu tiên, lựa chọn đối tượng cần quan sát trên màn Hình radar quan sát; hiển thị mục tiêu trên màn Hình điều khiển như trong Hình 4.9.

Các mục tiêu được phát hiện sẽ hiển thị trên màn Hình điều khiển, tại đây, nhân viên radar có thể quan sát các chuyển đổi, đo lường được tiêu cự và phương vị của mục tiêu. Để lấy được dạng sóng của các mục tiêu biển, Nghiên cứu sinh nhấn chuột phải để hiển thị ra một biểu mẫu hiển thị dạng sóng của tín hiệu phản xạ của mục tiêu.



(a) Dạng sóng tín hiệu phản xạ (b) Tệp dữ liệu mục tiêu

Hình 4.10: Dạng sóng của tín hiệu phản xạ và tệp dữ liệu.

Từ màn Hình hiển thị dạng sóng tín hiệu phản xạ như Hình 4.10a, Nghiên cứu sinh nhấn "export" sẽ đưa đến một giao diện Hình 4.10b để đặt tên tệp và lưu tín hiệu sóng phản xạ. Sóng phản xạ là một tín hiệu 500 mẫu được lưu trữ thành các tệp excel từ phần mềm trên thiết bị hiển thị. Hình 4.10a mô tả một ví dụ về dạng sóng thu được của một tàu và Hình 4.10b là kết quả các tệp excel thu được. Mỗi tệp tương ứng với một sóng phản hồi thu về. Dữ liệu thu thập được chia thành ba loại tàu gồm tàu cá,

tàu vận tải thường, tàu vận tải quân sự. Nghiên cứu sinh thu thập 611 mẫu tàu cá; 688 mẫu tàu vận tải dân dụng và 328 mẫu tàu vận tải quân sự.

4.2.6 Kết quả thí nghiệm với các phương pháp truyền thống

Bảng 4.1: Kết quả phân cụm dựa vào Kmeans và các đặc trưng.

Đặc trưng	chuẩn hoá	MI	A_MI	NorMI	A_RS	Complex	Flow
Raw	TRUE	14,41	17,48	17,61	03,05	24,78	51,59
Raw	FALSE	76,26	71,27	71,30	72,11	70,32	81,93
FFT	TRUE	49,11	46,22	46,28	46,96	46,00	65,86
FFT	FALSE	60,71	57,35	57,40	54,57	57,22	70,83
DCT	TRUE	60,81	59,58	59,63	47,23	61,75	67,72
DCT	FALSE	57,81	55,71	55,76	47,55	56,75	67,23
DWT	TRUE	44,63	45,79	45,86	34,65	50,05	61,95
DWT	FALSE	76,26	71,27	71,30	72,11	70,32	81,93

Trong phần này, các kết quả thí nghiệm dựa trên các thuật toán phân cụm truyền thống được trình bày. Thuật toán K-means được sử dụng với các đặc trưng khác nhau như FFT [25], DCT [26], DWT [27].Tín hiệu radar đã được lọc bằng phần mềm chuyên dụng, và chỉ trả về kết quả sau khi lọc. Các phép biến đổi FFT và DCT chỉ được sử dụng để biến đổi tín hiệu sang miền đặc trưng hỗ trợ các thuật toán AI.Riêng đối với phép biến đổi wavelet rời rạc (DWT), họ wavelet Haar đã được lựa chọn áp dụng với mức phân rã cấp 2 nhằm trích xuất chi tiết các đặc trưng tần số - thời gian của tín hiệu. Bên cạnh đó, các phương án chuẩn hoá dữ liệu cũng được áp dụng lên các đặc trưng để bù lại sự phi tuyến khi chuyển tín hiệu từ miền thời gian sang miền tần số.

Kết quả thí nghiệm được mô tả trong Bảng 4.1. Kết quả cho thấy đặc trưng trong miền thời gian cho kết quả tốt hơn đặc trưng fft hoặc dct. Điều này là vì các đặc trưng trong miền tần số thường được áp dụng để chống lại hiện tượng lệch pha trong miền thời gian; tuy nhiên tín hiệu radar nhận về là một tín hiệu dạng xung và ít bị ảnh hưởng của hiện tượng lệch thời gian lấy mẫu. Việc áp dụng đặc trưng DWT về cơ bản vẫn cho kết quả tốt như việc áp dụng tín hiệu trên miền thời gian. Nguyên nhân của việc này là vì bản chất của đặc trưng DWT là phân tách tín hiệu trong miền thời gian thành các thành phần tần số thấp và tần số cao nhưng vẫn giữ lại Hình dạng sóng trong miền thời gian. Các thành phần tần số cao đóng vai trò như nhiễu và các thành phần tần số thấp đóng vai trò như dạng sóng trong miền thời gian với tần số lấy mẫu thấp hơn. Chính vì thế kết quả phân cụm sử dụng đặc trưng DWT cho kết quả tương tự như phân cụm dựa trên tín hiệu trong miền thời gian.

Tuy nhiên nếu các đặc trưng được chuẩn hoá, thì kết quả phân cụm trên miền thời gian bị giảm đi rất

nhiều. Nguyên nhân của hiện tượng này là do các đặc trưng đều thể hiện cùng một đơn vị, do đó việc chuẩn hoá các đặc trưng không có ý nghĩa khi mô tả tín hiệu trong miền thời gian. Không những thế, việc chuẩn hoá đặc trưng còn làm tăng ảnh hưởng của những đặc trưng không quan trọng (thường xuất hiện ở đầu hoặc cuối dạng sóng phản hồi). Chính vì vậy, khi thực hiện chuẩn hoá đặc trưng thì kết quả phân cụm giảm đi rất nhiều xuống chỉ còn 14,41% đối với chỉ số MI. Ngược lại, khi mô tả tín hiệu trong miền dwt, việc phân tích tín hiệu thành các thành phần tần số cao cho phép rút trích các nhiễu và nhiễu này luôn có biên độ nhỏ. Do đó, chúng không bị ảnh hưởng khi chuẩn hoá dữ liệu. Tuy nhiên, điều này không làm thay đổi bản chất của việc chuẩn hoá các đặc trưng lên bản chất của tín hiệu. Do đó, chỉ số MI vẫn suy giảm đáng kể xuống 44,62% khi tiến hành chuẩn hoá đặc trưng từ DWT.

4.2.7 So sánh với các phương pháp học sâu tiên tiến

Phương pháp	MI	A_MI	NorMI	A_RS	Complex	Flow
AE [35]	77,28	72,19	72,22	73,03	71,19	82,51
VAE [36]	79,18	74,16	74,19	76,56	73,33	84,84
DEC [40]	77,59	72,47	72,50	73,44	71,46	82,27
DeepCluster [38]	75,33	70,39	70,43	87,12	71,75	69,45
Proposed method	84,37	80,01	80,03	84,98	80,07	90,39
Only attribute	28,48	26,72	26,81	19,92	26,60	26,61

Bảng 4.2: Kết quả phân cụm bằng kỹ thuật học sâu

Phần này so sánh phương pháp đề nghị với các phương pháp học sâu cho phân đoạn bao gồm AE [35], VAE [36], DEC [40] và DeepCluster [38]. Trong đó AE [35] và VAE [36] là hai phương pháp kinh điển mà khối rút trích đặc trưng và khối phân cụm dữ liệu được tách biệt với nhau. Điểm khác biệt lớn nhất giữa AE và VAE là phương pháp VAE sử dụng thêm ràng buộc phụ để các đặc trưng học được sẽ tuân theo phân phối Gaussian và nếu lấy mẫu một biến trong không gian đặc trưng thì sẽ tái tạo được một mẫu gần giống ở ngõ ra của khối giải mã. Tuy nhiên, ràng buộc này đôi khi không tốt cho việc phân cụm vì nó dễ làm các nhóm trộn lẫn với nhau. Chính vì lý do đó, kết quả ở Bảng 4.2.7 cho thấy phương pháp VAE cho kết quả không tốt bằng phương pháp AE. Giá trị MI của phương pháp AE là 77,28 % trong khi giá trị MI của phương pháp VAE là 79,18 %. So với các phương pháp truyền thống, đặc trưng của phương pháp AE có số chiều ít hơn, điều này đặc biệt hữu ích khi số lượng mẫu còn hạn chế. Chính vì vậy, việc nén thông tin giúp thuật toán phân loại hoạt động ổn định hơn.

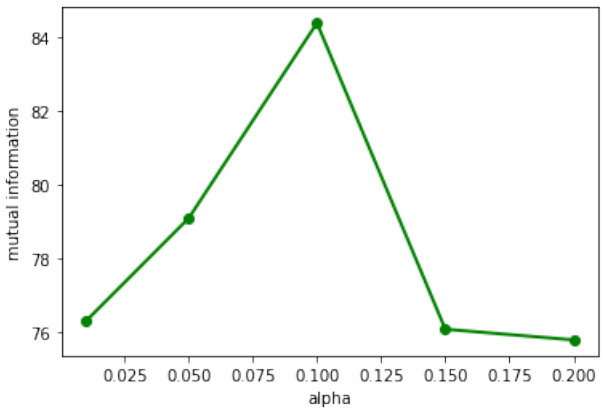
DeepCluster [38] sử dụng nhãn giả để huấn luyện khối rút trích đặc trưng. Điều này có một nhược

điểm là bộ rút trích đặc trưng phải rất ổn định để đảm các nhãn được gắn sai không làm hỏng bộ rút trích đặc trưng ban đầu. Tuy nhiên kết quả thí nghiệm cho thấy độ chính xác phân cụm bị suy giảm đáng kể. Điều này là vì ngay từ đầu số lượng các mẫu được phân cụm sai đã rất lớn. Như trong Hình 4.1, ta thấy nhóm 1 gồm rất nhiều loại tàu bị trộn chung với nhau bởi vì đặc điểm của biên độ đỉnh sóng. Khi số lượng các mẫu được phân cụm sai ngay từ khâu khởi tạo áp đảo số lượng các mẫu còn lại, thì việc dựa trên các nhãn giả để huấn luyện khối rút trích đặc trưng sẽ gây hiệu ứng tiêu cực.

DEC [38] là phương pháp phân cụm end-to-end mà ở đó phân phối của các mẫu thuộc cùng một nhóm trên miền đặc trưng sẽ có tính phân cụm tốt hơn. Tức là các đặc trưng sẽ được học để khoảng cách của một điểm trong miền đặc trưng tới tâm của nhóm sẽ nhỏ hơn. Tuy nhiên điều này không thể thay đổi quá nhiều kết quả phân cụm ban đầu vì sau mỗi lần cập nhập thì các mẫu thuộc cùng một nhóm cũng rất khó thay đổi nhóm của nó. Tuy kết quả được tăng cường không nhiều, nhưng kỹ thuật này cũng giúp cải thiện giá trị MI lên được 0,31 điểm % so với phương pháp AE, và cải thiện 1,39 điểm % so với các phương pháp không dựa trên học sâu.

So với phương pháp AE [35], các phương pháp DeepCluster [38] hoặc DEC [40] phức tạp hơn và đòi hỏi nhiều dữ liệu cũng như các mô hình được huấn luyện trước để có thể đạt kết quả tốt. Tuy nhiên các điều kiện này vẫn chưa đạt được khi dữ liệu huấn luyện còn ít. Do đó, trong trường hợp này, lựa chọn AE [35] làm phương pháp phân cụm sẽ có lợi hơn.

Khác với các phương pháp học sâu kinh điển, phương pháp được đề nghị có sự hỗ trợ của thông tin thuộc tính để cải thiện bộ rút trích đặc trưng. Trong bối cảnh thiếu hụt dữ liệu mẫu, việc tích hợp kiến thức chuyên gia dưới dạng các nhãn thuộc tính Hình thái học nhằm ép buộc bộ rút trích đặc trưng phải tuân theo các quy luật vật lý của tín hiệu radar, từ đó tăng cường độ hội tụ và tính chính xác của mô hình, là một giải pháp hoàn toàn phù hợp. Cụ thể, kiến thức chuyên gia được sử dụng ở đây chính là sự hiểu biết về các đặc tính vật lý và cấu trúc Hình học của từng loại tàu (tàu cá, tàu quân sự, tàu vận tải) phản ánh trực tiếp lên Hình thái của sóng radar phản hồi. Những hiểu biết này đã được cụ thể hóa thành 6 thuộc tính đặc trưng (bao gồm: số lượng đỉnh, biên độ và độ dao động của phần đỉnh cũng như phần đáy sóng) như đã được định nghĩa chi tiết tại Mục 4.2.2. Để tích hợp vào mô hình, các kiến thức này được lượng hóa thành các vector thuộc tính (nhãn a) và gán cho từng mẫu dữ liệu. Trong quá trình huấn luyện mô hình rút trích đặc trưng (Encoder), bên cạnh nhiệm vụ tái tạo lại tín hiệu gốc thông thường (L_R), mô hình được yêu cầu phải đồng thời tối ưu hóa hàm mục tiêu dự đoán thuộc tính (L_A). Sự ràng buộc đa nhiệm này ép buộc mạng nơ-ron phải học cách đối chiếu không gian đặc trưng ẩn với các quy luật vật lý do chuyên gia đúc kết, từ đó giúp bộ trích xuất đặc



Hình 4.11: Ảnh hưởng của hệ số α tới chỉ số mutual information.

trưng hội tụ chính xác và tạo ra các biểu diễn mạnh mẽ ngay cả khi số lượng mẫu dữ liệu không đủ lớn. Theo đó, bộ rút trích đặc trưng không chỉ giúp tái tạo lại thông tin ban đầu mà đôi khi còn hữu ích trong việc dự đoán các thuộc tính của một mẫu. Ta có thể thấy giá trị MI được tăng đáng kể lên 84,13 % ứng với $\alpha = 0{,}1$. Cần lưu ý là nếu chỉ tiến hành phân cụm dựa vào thông tin của thuộc tính, thì giá trị phân cụm cho các chỉ số MI, A_MI, N_MI, RS, A_RS, CS, và FM_S sẽ lần lượt như sau 0,2848,0,2672,0,2680,0,5083,0,1992,0,2660, 0,4838. Qua có có thể thấy việc gắn atrribute không thực sự giúp ích trực tiếp cho việc phân cụm; nhưng khi kết hợp với hàm mục tiêu tái tạo tín hiệu, thì các đặc trưng học được sẽ hỗ trợ tìm được các đặc trưng tốt hơn cho việc phân cụm. Chính vì lý do này mà khi thiết kế khối phân cụm, ta cần sử dụng nhiều hơn một lớp mạng neuron để giảm bớt ảnh hưởng từ việc dự đoán attribute đối với các đặc trưng dùng cho phân cụm.

Ta có thể thấy rằng nếu tăng giá trị α đủ lớn, thì kết qủa MI không còn tăng nữa mà có xu hướng giảm như Hình 4.11. Điều này là vì các thuộc tính được gán theo nhóm không thực sự hữu ích cho việc phân cụm. Nếu quá tin tưởng vào các thuộc tính này thì kết quả phân cụm sẽ suy giảm hơn.

4.3 mô hình và đánh giá

Khác với bài toán phân nhóm, bài toán phân loại được bổ sung thêm các nhãn cho từng loại tàu. Để có thể nhận biết các loại tàu khác nhau, nhân viên ngành radar thường dựa vào kinh nghiệm quan sát sóng phản xạ. Các nhân viên này phải có kinh nghiệm lâu năm. Đồng thời, họ phải kết hợp việc sử dụng kính ngắm quang học và camera quan sát, nhưng chỉ có thể nhận biết chính xác khi tàu ở cự ly rất gần và trong "vùng mù" của radar. Khi tàu ở xa, hoặc hoạt động trong môi trường nhiễu, khi hướng tàu, tốc độ tàu thay đổi, điều kiện thời tiết thay đổi, và sóng biển lớn, vấn đề nhận biết trở nên rất khó

khăn với nhân viên radar.

Tín hiệu phản xạ sau khi thu được không chỉ chứa đựng thông tin cần thiết như mục tiêu lớn với tín hiệu phản xạ trở về lớn, mục tiêu nhỏ với tín hiệu phản xạ trở về nhỏ; cùng một mục tiêu nhưng ở xa, tín hiệu phản xạ về nhỏ, ở gần, tín hiệu phản xạ về lớn. Bề mặt mục tiêu phẳng, ổn định, tín hiệu phản xạ về ổn định; bề mặt mục tiêu nhấp nhô, biên độ tín hiệu phản xạ về thay đổi, không ổn định. Ngoài ra, nó cũng mang rất nhiều tín hiệu nhiễu, nguyên nhân có thể là do điều kiện thời tiết, do cấp độ sóng biển, do góc quan sát, do cự ly hoạt động của mục tiêu, do thiết bị radar hoặc những yếu tố khác.

Các nhà khoa học nghiên cứu và phân tích để tăng cường việc đảm bảo an toàn hàng hải và an ninh quốc gia. Xử lý tín hiệu phản xạ từ mục tiêu trở về từ đó nhận dạng được kiểu loại trước những thay đổi về điều kiện thời tiết, chất liệu của mục tiêu, cự ly và phương vị của mục tiêu thay đổi... thực sự là một thử thách lớn đối với các nhà khoa học trong việc phân loại.

Nhằm giúp nhân viên trạm radar làm việc tốt hơn, một giải pháp đơn giản là huấn luyện một mô hình trí tuệ nhân tạo để lưu trữ kiến thức của các chuyên gia phân tích tín hiệu radar và hỗ trợ nhân viên trực trạm. Về mặt tổng quát, một hệ thống phân loại tín hiệu được mô tả như Hình 4.12. Đầu tiên, tín hiệu sẽ được tiền xử lý để lọc nhiễu, sau đó được rút trích đặc trưng để đưa ra các thông tin phù hợp. Các đặc trưng đó sẽ được đưa vào các bộ phân loại, từ đó đưa ra kết quả phân loại.

Hình 4.12: Sơ đồ xử lý tín hiệu phản xạ từ mục tiêu

4.3.1 Rút trích đặc trưng

Sự thành công của một thuật toán học máy phụ thuộc rất lớn vào các đặc trưng được lựa chọn. Nếu các đặc trưng được lựa chọn có độ ổn định cao đối với sự thay đổi của môi trường, độ chính xác của bộ phân loại sẽ được cải thiện đáng kể. Một cách truyền thống, các tín hiệu 1D trong miền thời gian, như sóng phản xạ của radar, sẽ bị ảnh hưởng đáng kể bởi việc lấy mẫu vùng dữ liệu cần được nhận dạng. Chính vì vậy, các đặc trưng trong miền tần số như Fourier Transform [70] thường được sử dụng trong các bài toán phân loại tín hiệu 1D.

Không giống như các trường hợp thông thường, sóng radar xung được điều chế bởi một xung vuông

và tín hiệu thu về là một mảng với kích thước không đổi tương ứng với bề rộng của xung điều chế. Chính vì vậy, ứng dụng của Nghiên cứu sinh không tồn tại những thách thức về việc xác định vùng dữ liệu cần nhận dạng. Như trong Hình 4.4, 4.5, và 4.6 tín hiệu thu về luôn luôn là một mảng 500 giá trị trong miền thời gian. Hơn thế nữa, các tín hiệu này khá ổn định nếu các tàu di chuyển theo cùng một hướng. Do đó, việc sử dụng tín hiệu trong miền thời gian để thực thi các thuật toán học máy cũng là một giải pháp có cơ sở.

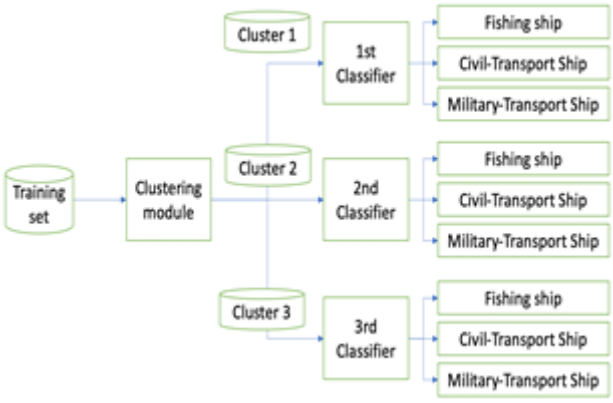
4.3.2 Các mô hình phân loại

Đối với bài toán phân loại, các kỹ thuật Máy Vector Hỗ trợ (SVM) [71, 72] và Mạng Nơ-ron (Nơ-ron) [73, 74] là hai phương pháp nổi tiếng có thể giải quyết các thách thức từ tập dữ liệu phi tuyến. Trong khi mạng nơ-ron dựa vào các hàm kích hoạt phi tuyến và các lớp ẩn để thể hiện ánh xạ phi tuyến tính, bộ phân loại SVM mô tả sự phi tuyến dựa vào các hạt nhân (kernel) được xác định trước. Thông thường, chất lượng của bộ phân loại phụ thuộc vào mức độ phù hợp giữa độ phức tạp của mô hình và tập dữ liệu huấn luyện. Nếu mô hình cực kỳ phức tạp nhưng tập dữ liệu lại đơn giản, chúng ta có thể gặp hiện tượng quá khớp (overfitting). Ngược lại, nếu mô hình đơn giản nhưng bộ dữ liệu là một bộ dữ liệu phức tạp, mô hình có thể gặp hiện tượng thiếu khớp (Underfitting). Việc tinh chỉnh mô hình, về mặt bản chất, là lựa chọn độ phức tạp của mô hình để phù hợp với độ phức tạp của tập dữ liệu. Trong trường hợp này, độ chính xác trong quá trình huấn luyện sẽ gần bằng hoặc cao hơn một chút so với độ chính xác trong quá trình đánh giá; và mô hình được coi là đủ tổng quát. Trong khi Nơ-ron dựa vào số lớp và số nút trong mỗi lớp ẩn để xác định độ phức tạp của một mô hình; SVM dựa vào băng thông kernel để kiểm soát độ phức tạp của nó. Chọn các tham số điều khiển phù hợp là một nhiệm vụ quyết định để huấn luyện thành công bộ phân loại.

4.3.3 Bộ phân loại dựa trên học sâu

Những năm gần đây, kỹ thuật học sâu đã phát triển mạnh mẽ để có thể vừa học các đặc trưng vừa huấn luyện mô hình phân loại trong một quy trình huấn luyện duy nhất. Ưu điểm của quá trình này là giảm sự can thiệp của con người vào quá trình huấn luyện, các đặc trưng sẽ được tự học để phù hợp với dữ liệu và tác vụ cần thực thi. Tuy nhiên sự thành công của phương pháp này chỉ có thể đạt được khi dữ liệu huấn luyện đủ lớn. Nếu dữ liệu huấn luyện nhỏ, các phương pháp truyền thống sẽ có thể đạt được độ chính xác cao hơn.

4.3.4 Tổng quan hệ thống



Hình 4.13: Tổng quan hệ thống

Do đặc thù bộ dữ liệu thu được vẫn còn ít, trong tiểu luận này, Nghiên cứu sinh chỉ sử dụng các phương pháp truyền thống để thực hiện việc phân loại tàu.

Trong chương này, Nghiên cứu sinh giới thiệu chi tiết về phương pháp được đề xuất. Đối với một tập dữ liệu nhất định, Nghiên cứu sinh sử dụng thuật toán K-means để tách tập dữ liệu này thành ba cụm. Đối với mỗi nhóm, một Nơ-ron được sử dụng để phân loại các loại tàu. Các loại bao gồm tàu cá, tàu vận tải dân dụng và tàu vận tải quân sự. Trong giai đoạn thử nghiệm, một mẫu thử nghiệm được chỉ định cho một cụm cụ thể. Tùy theo cụm mà mẫu thử nghiệm được chỉ định, một bộ phân loại tương ứng sẽ được chọn để phân loại. Trong tiểu mục tiếp theo, phân cụm K-means và mạng Neuron sẽ được giới thiệu chi tiết.

4.3.5 Phân cụm đối tượng Kmeans

Thuật toán phân cụm là một thuật toán chia một tập hợp các mẫu đầu vào thành các cụm khác nhau. Cụm là một nhóm các điểm dữ liệu tương tự nhau dựa trên mối quan hệ của chúng với các điểm dữ liệu xung quanh. Một trong những thuật toán phân cụm đơn giản và dễ thực hiện nhất là kỹ thuật K-means.

Thuật toán K-means được sử dụng tốt nhất trên các tập dữ liệu nhỏ hơn vì nó lặp lại trên tất cả các điểm dữ liệu. Điều đó có nghĩa là sẽ mất nhiều thời gian hơn để phân loại các điểm dữ liệu nếu có một lượng lớn chúng trong tập dữ liệu. Với tập dữ liệu $X = [x_1, x_2, , , , , x_N] \in R^{dx_N}$ bao gồm N mẫu được biểu thị bằng d đặc trưng, thuật toán K-Means sẽ phân tách tập dữ liệu này thành K cụm. Trong mỗi cụm, các thành viên của nó phải càng giống nhau càng tốt. Mỗi cụm được đại diện bởi một tâm

cụm là $m_k \in R^{dx_1}(k = 1 \sim K)$; và đối với mỗi điểm trong tập dữ liệu, một nhãn phải được gán để xác định mẫu thuộc về cụm nào. Thông thường, nhãn có thể được biểu thị bằng một vectơ một chiều là $y_{ik} \in \{0,1\}$ và $\sum_{(k=1)}^{K} y_{ik} = 1$. Nếu $y_{ik} = 1.$, có nghĩa là mẫu thứ i^{th} thuộc cụm thứ k^{th}. Ở đây, Nghiên cứu sinh hy vọng rằng một mẫu sẽ gần với trung tâm cụm mà nhãn của nó thuộc về. Do đó, hàm mất trong phương trình (4.9) được sử dụng để tìm hiểu $\{y_i, m_k\}$.

$$J = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,j} \|x_i - m_k\|_2^2 \tag{4.9}$$

Kí hiệu $Y = [y_1, y_2, ,,,, y_N]$ và $M = [m_1, m_2, ,,,, m_K]$ là dạng ma trận của tất cả các nhãn và trung tâm cụm, giải pháp Y và M có thể được tìm thấy bởi phương trình (4.10).

$$\begin{gathered} Y, M = \underbrace{\text{agrmin}}_{Y,M} \sum_{i=1}^{N} \sum_{k=1}^{K} y_{ik} \|x_i - m_k\|_2^2 \\ \text{Điều kiện: } \mathrm{y_{ik}} \in \{0,1\} \quad \forall \mathrm{i,k}; \sum_{j=1}^{K} y_{ij} = 1 \forall i \end{gathered} \tag{4.10}$$

Vì Y và M thuộc hai nhóm tham số khác nhau nên thuật toán Expectation-Maximization [75] được áp dụng để giải phương trình (4.10). Ở đây, M được khởi tạo ngẫu nhiên M và Y được lựa chọn để hàm mục tiêu J đạt giá trị nhỏ nhất. Sau đó, nghiệm của Y có thể được ước tính bằng phương trình (4.11):

$$\begin{gathered} y_i = \arg\min_{y_i} \sum_{k=1}^{K} y_{ik} \|x_i - m_k\|_2^2 \\ \text{Điều kiện: } y_{ik} \in \{0,1\} \forall j; \sum_{k=1}^{K} y_{ik} = 1 \end{gathered} \tag{4.11}$$

Sau đó chúng ta cố định Y và ước lượng M theo công thức (4.12)

$$m_j = \frac{\sum_{i=1}^{N} y_{ik} x_i}{\sum_{i=1}^{N} y_{ik}} \tag{4.12}$$

Để tìm Y và M tối ưu, phương trình (4.11) và (4.12) được lặp lại cho đến khi các tham số được chuyển đổi

4.4 Một số kết quả thí nghiệm

4.4.1 So sánh các đặc trưng trong miền tần số và trong miền thời gian

Một cách truyền thống, các bài toán phân loại tín hiệu 1D [76],[77] sẽ phải phân đoạn tín hiệu trước khi tiến hành phân loại. Bởi vì tín hiệu trong miền thời gian sẽ dễ bị ảnh hưởng bởi việc phân đoạn, các đặc trưng tần số của tín hiêu 1D sẽ được rút trích để loại bỏ ảnh hưởng này. Mặc dù vậy, trong ứng dụng này, sóng radar phát ra được điều chế bởi một xung vuông nên bài toán phân đoạn đã được giải quyết tốt. Do đó, tín hiệu trong miền thời gian cũng có thể là một tín hiệu đáng tin cậy để thực hiện các thuật toán học máy. Trong thí nghiệm này, Nghiên cứu sinh khảo sát ảnh hưởng của các đặc trưng trong miền tần số và trong miền không gian đối với các thuật toán học máy. Hai bài toán học máy cơ bản là phân nhóm và phân loại sẽ được áp dụng với các đặc trưng FFT và đặc trưng trong miền thời gian. Ứng với mỗi bài toán, Nghiên cứu sinh sử sụng 33% dữ liệu để kiểm tra và 66% dữ liệu cho huấn luyện. Các thuật toán được thực thi dựa trên thư viện sklearn và tham số random seed được cố định để đảm bảo việc lựa chọn tập huấn luyện và tập đánh giá là giống nhau qua mỗi bước thí nghiệm.

Đối với bài toán phân nhóm, Nghiên cứu sinh sử dụng thuật toán K-means để phân chia dữ liệu thành 3 nhóm. Kết quả phân nhóm được trình bày trong Bảng 4.3. Ở đây, Nghiên cứu sinh sử dụng biến đổi FFT để rút trích đặc trưng trong miền tần số. Nghiên cứu sinh so sánh đặc trưng trong miền tần số này với kết quả trong miền không gian. Bên cạnh đó, Nghiên cứu sinh áp dụng kỹ thuật chuẩn hoá dữ liệu (loại trừ trị trung bình và độ lệch chuẩn) và lấy logarit biên độ của tín hiệu tần số để bù lại sự phi tuyến gây ra bởi biến đổi FFT. Nghiên cứu sinh thống kê số lượng tàu cá, tàu vận tải quân sự và tàu vận tải thường trong mỗi nhóm. Kết quả thí nghiệm trong Bảng 4.3 cho thấy kết quả phân cụm trên miền thời gian tốt hơn nhiều kết quả phân cụm trên miền tần số. Cụ thể, trong miền thời gian và chỉ xét tới tập huấn luyện, nhóm C0 chỉ chứa hoàn toàn các tàu vận tải thường, nhóm C1 có 92.3% là tàu cá, nhóm C2 chỉ chứa tàu vận tải chứ không có tàu cá nào (với 69% là tàu vận tải quân sự). Hơn thế nữa, kết quả phân nhóm cho tập test chứng minh rằng việc phân nhóm trên tập test là phù hợp với kết quả phân nhóm trên tập huấn luyện. Trong quá trình huấn luyện, nếu một loại tàu không xuất hiện trong một nhóm, thì trong tập test loại tàu đó cũng không xuất hiện trong nhóm tương ứng. Điều này cho thấy việc phân nhóm sẽ không giới thiệu bất kỳ lỗi nào trong quá trình phân loại sau này.

Ngược lại, nếu chúng ta sử dụng FFT feature để phân nhóm, ta thấy kết quả phân nhóm rất hỗn loạn. Các nhóm không tập trung vào bất kỳ loại tàu nào. Khi sử dụng kỹ thuật chuẩn hoá dữ liệu hoặc biến

đổi log để bù lại sự phi tuyến trong tín hiệu, kết quả phân nhóm được cải thiện theo hướng tập trung vào một loại tàu cụ thể cho mỗi nhóm. Mặc dù vậy, độ phân tách dữ liệu vẫn không thể so sánh được khi thực hiện phân nhóm trong miền không gian. Điều này minh chứng cho việc tín hiệu trong miền tần số không thực sự giúp ích cho ứng dụng phân loại tàu trên biển vì tín hiệu radar phản hồi không bị ảnh hưởng bởi bài toán segment tín hiệu. Bên cạnh đó, phép biến đổi FFT còn giới thiệu những đặc tính phi tuyến vốn gây khó khăn cho các thuật toán học máy.

Bảng 4.3: Kết quả phân loại với các đặc tính khác nhau của các loại tàu.

Quá trình huấn luyện				
		Tàu cá	Tàu vận tải quân sự	Tàu vận tải thường
FFT	C0	418	14	21
	C1	0	13	205
	C2	1	186	232
FFT + Scaling	C0	0	18	299
	C1	399	33	28
	C2	20	162	201
FFT + Log	C0	367	9	9
	C1	2	45	308
	C2	50	159	141
Time domain	C0	0	0	341
	C1	419	7	28
	C2	0	206	89
Quá trình kiểm tra đánh giá				
FFT	C0	192	6	10
	C1	0	9	114
	C2	0	100	106
FFT + Scaling	C0	0	16	116
	C1	186	15	12
	C2	6	84	102
FFT + Log	C0	175	2	6
	C1	1	32	160
	C2	16	81	84
Time domain	C0	0	0	180
	C1	192	4	14
	C2	0	111	36

Trong quá trình đánh giá, đưa vào một mẫu để kiểm tra, bộ phân nhóm ở trên sẽ được sử dụng để đưa dữ liệu kiểm tra vào một trong ba nhóm đã được tạo ra trong quá trình đào tạo. thuộc nhóm mà bài kiểm tra mẫu được chỉ định, bộ phân loại.

Bảng 4.4: Tóm tắt các cấu Hình mạng Nơ-ron và thông số huấn luyện

Parameter	Solver	Hệ số Regularization	Hidden Layer Size	Random State
S1	lbfgs	1e-5	(300,200,100,50)	100
S2	lbfgs	1e-5	(300,100,50)	100
S3	lbfgs	1e-5	(200,75)	100
S4	lbfgs	1e-5	(100,75)	100
S5	lbfgs	1e-5	(75,50)	100

Không chỉ thực hiện phân nhóm bằng các đặc trưng khác nhau, Nghiên cứu sinh thực hiện việc phân loại tàu trực tiếp dựa trên các đặc trưng đã được sử dụng trong bài toán phân nhóm. Nghiên cứu sinh sử dụng ba cấu Hình mạng Nơ-ron (S1,S3,S5) để đánh giá vai trò của các đặc trưng. Trong các bộ phân loại này, cấu Hình S1 có độ phức tạp cao nhất vì gồm nhiều lớp và số nút ở mỗi lớp đều cao. mô hình S5 có độ phức tạp thấp nhất vì chỉ có 2 lớp ẩn và số lượng nút cho mỗi lớp chỉ đều thấp. Chi tiết của các bộ phân loại này được mô tả trong Bảng 4.4. Kết quả thí nghiệm trong Bảng 4.5 cho thấy đặc trưng FFT không thích hợp để thực hiện việc phân loại tàu. Kết quả phân loại rất thấp chứng tỏ mô hình không thể hội tụ tốt để phân loại tàu. Nếu tiến hành bù lại các yếu tố phi tuyến bằng chuẩn hoá dữ liệu [78] và hàm log, mô hình có thể học nhưng độ chính xác không cao. Giá trị tốt nhất có thể đạt được trên tập test là 96% trong khi độ chính xác khi phân loại trên miền thời gian là 98%. Kết quả này thống nhất với kết quả khi thực hiện thuật toán phân nhóm khi chỉ ra rằng tín hiệu trong miền thời gian là đủ tốt để xử lý tín hiệu radar trên biển.

Bảng 4.5: Kết quả phân loại bằng các đặc trưng khác nhau với tỷ lệ huấn luyện/kiểm tra là 2.

Pha test			
		ACC	F1-Score
FFT	S1	0,1195	0,2364
	S3	0,5680	0,6312
	S5	0,3910	0,4357
FFT + Scaling	S1	0,9606	0,9608
	S3	0,9606	0,9608
	S5	0,9492	0,9497
FFT + Log	S1	0,9414	0,9422
	S3	0,9243	0,9255
	S5	0,0945	0,9255
Time domain	S1	0,9813	0,9814
	S3	0,9795	0,9795
	S5	0,9795	0,9795

4.4.2 Ảnh hưởng của thuật toán phân nhóm lên bài toán phân loại

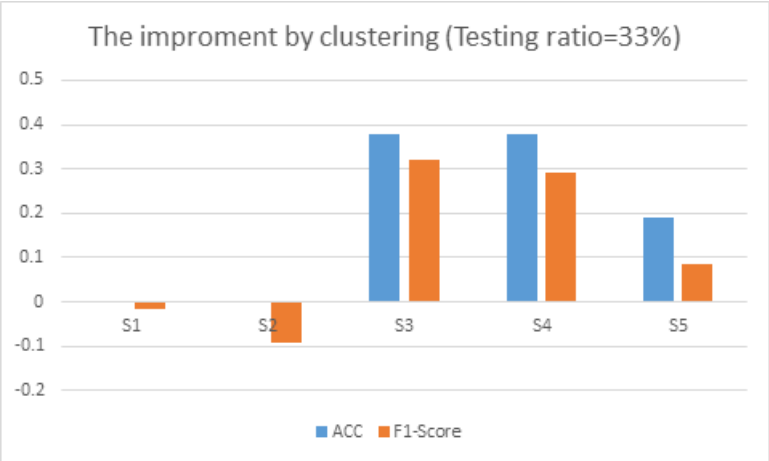
Trong phần này, Nghiên cứu sinh chứng minh tính hiệu quả của phương pháp được đề nghị khi sử dụng thuật toán phân nhóm để tiền xử lý dữ liệu trước khi đưa vào bài toán phân loại. Giống như các bài toán phân loại truyền thống, Nghiên cứu sinh sử dụng accuracy và F1-score để đánh giá hiệu quả của hệ thống. Mặc dù vậy, để nhấn mạnh sự đóng góp của quá trình phân nhóm, Nghiên cứu sinh sử dụng tỷ lệ tăng cường của ACC và F1-score như trong phương trình (4.13).

$$E = \frac{P_{clus} - P}{P} 100 \tag{4.13}$$

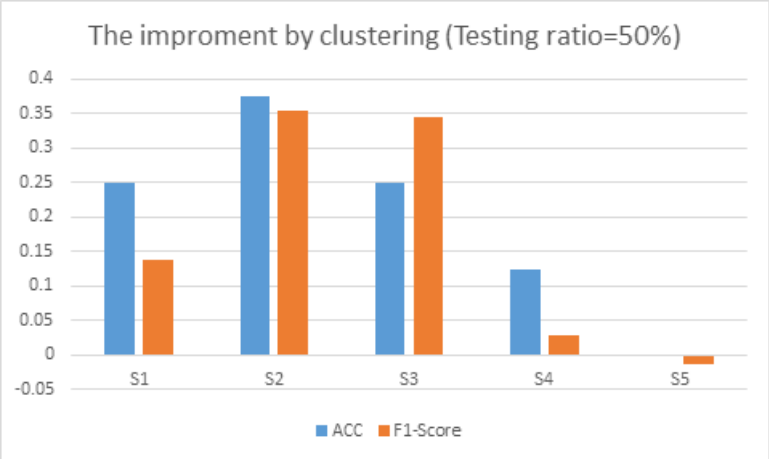
Ở đây, P là hiệu năng của hệ thống khi không có quá trình tiền xử lý phân nhóm; P_{clus} là hiệu năng của hệ thống khi có quá trình tiền xử lý phân nhóm. Hiệu năng ở đây bao gồm độ chính xác (ACC) hoặc điểm F1 (F1-score). Nếu giá trị này lớn hơn 0, ta có thể kết luận phép tiền xử lý bằng cách phân nhóm giúp cải thiện chất lượng của việc phân loại. Ngược lại, nếu giá trị này nhỏ hơn 0, có nghĩa là phép phân nhóm không giúp tăng cường perfomrnace. Nghiên cứu sinh tiến hành thí nghiệm với nhiều cấu Hình Nơ-ron khác nhau từ rất dày đặc (dense) (S1) tới rất thưa thớt (sparse)(S5). Chi tiết của các cấu Hình Nơ-ron và các thông số setting được liệt kê như trong Bảng 4.4.

Hình 4.14 thể hiện mức độ tăng cường của độ chính xác và F1-Score khi lựa chọn tỷ lệ huấn luyện/testing là 66%/33%. Kết quả cho thấy khi mạng Nơ-ron rất dày đặc, việc phân cụm không giúp ích được nhiều mà thậm chí còn làm giảm hiệu suất. Điều này xảy ra bởi vì khi thực hiện phân cụm, lượng dữ liệu cho các phần phân loại thành phần sẽ giảm đi. Do đó, các bộ phân loại quá dày đặc sẽ bị quá tải vì không đủ dữ liệu để học. Ngược lại, khi không sử dụng phân cụm, bộ phân loại sẽ có nhiều dữ liệu hơn để học. Vì vậy với những mạng Nơ-ron dày đặc thì việc tiền xử lý bằng cách phân cụm không giúp tăng cường hiệu suất. Như trong cấu Hình S1, S2, độ chính xác được giữ nguyên và F1-Score bị giảm nhẹ. Ngược lại, với những mạng Nơ-ron thưa thớt như cấu Hình S3,S4 ta có thể thấy phép toán phân cụm giúp tăng cường hiệu suất một cách rõ ràng trong cả ACC và F1-Score. Khi cấu Hình mạng trở nên thưa thớt, một lượng dữ liệu nhỏ có thể giúp học tốt mô hình. Môt Hình S5 là một mạng Nơ-ron rất nhỏ. Chính vì vậy nó có xu hướng xảy ra thiếu phù hợp trên cả hai phương pháp có và không có cụm. Trong tất cả các cấu Hình này, giá trị ACC lớn nhất khi không sử dụng phân cụm là 98% trong khi giá trị ACC lớn nhất khi sử dụng phân cụm là 99%. Điều này chứng minh rằng tiền xử lý bằng cách phân cụm có thể giúp cải thiện hiệu suất tốt hơn ở mức tổng thể.

Hơn thế nữa, khi giảm số lượng các mẫu huấn luyện xuống, Nghiên cứu sinh nhận ra rằng khả năng



Hình 4.14: Độ tăng cường của ACC và F1-Score khi tỷ lệ huấn luyện/testing là 66%/33%



Hình 4.15: Độ tăng cường của ACC và F1-Score khi tỷ lệ huấn luyện/testing là 50%/50%

xảy ra quá phù hợp sẽ cao hơn nếu không sử dụng kỹ thuật phân nhóm. Như trong Hình 4.15, Nghiên cứu sinh tăng tỷ lệ kiểm tra lên 50%. Lúc này, việc phân nhóm giúp tăng hiệu năng trên cả các cấu Hình S1-S2 chứ không chỉ trong các cấu Hình S3-S4. Điều này là vì khi tăng lượng dữ liệu kiểm tra, lượng dữ liệu huấn luyện sẽ giảm. Khi đó, lượng dữ liệu được sử dụng cho việc huấn luyện cấu hinh S1, S2 sẽ không còn đủ nữa và sẽ xảy ra hiện tượng quá khớp (overfitting). Bằng cách sử dụng phân nhóm, các dữ liệu khó đã được phân cụm vào cùng một nhóm. Như thể hiện trong Bảng 4.3, đa số các mẫu khó đều tập hợp trong nhóm C1. Vì vậy, ngay cả khi sử dụng ít dữ liệu hơn, số lượng mẫu dùng để huấn luyện cho nhóm C1 vẫn không bị suy giảm đáng kể. Ví dụ, theo Bảng 4.3, nếu 66% dữ liệu được dùng cho training, ta sẽ có 419 mẫu là tàu cá, 7 mẫu là tàu vận tải quân sự, và 28 mẫu là tàu vận tải thưởng. Khi đó việc giảm số lượng mẫu huấn luyện sẽ tập trung vào giảm số lượng mẫu tàu cá; các tàu vận tải vẫn sẽ giữ lại được số mẫu gần giống ban đầu. Do đó kết quả huấn luyện không bị ảnh hưởng nhiều. Ở một khía cạnh khác, việc giảm số lượng mẫu huấn luyện cho các nhóm dễ (C0-C2)

thực ra không gây ảnh hưởng tới chất lượng hệ thống. Nguyên nhân của hiện tượng này là vì các mẫu này đã là các mẫu dễ nên không gây ra hiện tượng quá .

Nghiên cứu sinh đề xuất một phương pháp dựa trên phân nhóm để phân loại các loại tàu trên mặt biển. Kết quả phân nhóm các tín hiệu trong miền thời gian chứng tỏ rằng quá trình phân nhóm sẽ không gây ra bất kỳ ảnh hưởng xấu nào trong quá trình phân loại. Hơn thế nữa, kết quả thí nghiệm với nhiều cấu Hình Nơ-ron cho thấy sử dụng phân nhóm sẽ tăng 1 điểm % trên độ chính xác của bộ phân loại. Cuối cùng, khi giảm số lượng mẫu để huấn luyện, việc áp dụng phân nhóm cho bài toán phân loại tàu sẽ làm giảm khả năng bị quá phù hợp của bộ phân loại.

4.5 Kết luận của chương

Trong chương này, Nghiên cứu sinh đã nghiên cứu giải pháp để nâng cao khả năng phát hiện các loại tàu biển từ cảm biến radar. Bộ dữ liệu được sử dụng là bộ dữ liệu được thu thập thực thế từ các đài radar xung của Việt Nam. Các đặc trưng dựa trên kinh nghiệm của chuyên gia đã được giới thiệu để mô tả các đặc điểm của sóng radar phản hồi trên vùng biển Việt Nam ứng với các loại tàu thường gặp như tàu cá, tàu hải cảnh, tàu hàng. Dựa trên các đặc trưng này, một mô hình phân cụm học sâu dựa trên sự hướng dẫn của kiến thức chuyên gia đã được giới thiệu để đánh giá chất lượng của các đặc trưng chuyên gia. Sau đó, một mô hình phân loại dựa trên phân cụm đã được đề xuất để phân loại tín hiệu radar tàu biển. Các đóng góp chính của chương này có thể được liệt kê như sau:

- Xây dựng bộ dữ liệu sóng phản hồi radar thực thế cho phân loại tàu biển. Trong lĩnh vực viễn thông, dữ liệu thực tế là rất khó thu thập để nghiên cứu. Nghiên cứu này thu thập dữ liệu thực từ các trạm quan trắc của Việt Nam để tiến hành phân tích không chỉ bằng các phương pháp học máy mà còn dựa trên kiến thức chuyên sâu của nhân viên đài radar.
- Xây dựng thuật toán phân cụm dựa theo sự hướng dẫn của các đặc trưng chuyên gia. Kết quả cho thấy các đặc trưng này giúp cải thiện 7,09% chỉ số MI so với phương pháp truyền thống khi chưa sử dụng đặc trưng hướng dẫn từ chuyên gia.
- Dựa trên kết quả phân cụm, một mô hình phân loại đã được đề xuất sử dụng mạng noron và kết quả đạt được độ chính xác lớn hơn 98% trên bộ dữ liệu được thử nghiệm.

Mặc dù phương pháp phân loại dựa trên tín hiệu radar đã đạt được độ chính xác cao (trên 99% trên tập kiểm thử), bản chất của tín hiệu radar vẫn tồn tại những hạn chế cố hữu về mặt định danh chi tiết.

Radar có thể phát hiện sự hiện diện và phân nhóm tàu dựa trên kích thước phản xạ (RCS), nhưng khó có thể xác định chính xác số hiệu tàu, màu sắc, hoặc các đặc điểm nhận dạng trực quan khác. Để nâng cao độ tin cậy và khả năng giám sát toàn diện, thông tin từ radar cần được sử dụng như một tín hiệu dẫn đường (cueing signal) để kích hoạt và định hướng cho hệ thống quan sát quang học. Đặc biệt, tại các khu vực có mật độ tàu thuyền đông đúc, sự kết hợp này sẽ giúp hệ thống bóc tách rõ ràng từng chiếc tàu nằm sát nhau, khắc phục triệt để tình trạng nhiễu và chồng lấn tín hiệu mà radar thường hay gặp phải.

Chương 5

KẾT LUẬN VÀ KIẾN NGHỊ

Chương 5 tổng kết các kết quả đạt được của luận án, nhấn mạnh những đóng góp mới về mặt phương pháp và thực nghiệm trong bài toán nhận dạng, phân loại tàu biển dựa trên radar và camera. Các cải tiến về mô hình học sâu và cách tiếp cận kết hợp tri thức chuyên gia được khẳng định thông qua các kết quả định lượng. Bên cạnh đó, chương cũng phân tích những hạn chế còn tồn tại và đề xuất các hướng nghiên cứu tiếp theo, bao gồm mở rộng dữ liệu đa cảm biến, triển khai thực địa và tối ưu hóa hệ thống theo thời gian thực. Những định hướng này tạo tiền đề cho việc phát triển các hệ thống cảnh giới bờ biển thông minh trong tương lai.

5.1 Kết luận

Giám sát và bảo đảm an ninh, an toàn trên biển và vùng ven bờ là một nhiệm vụ có ý nghĩa đặc biệt quan trọng đối với phát triển kinh tế biển, bảo vệ chủ quyền quốc gia và duy trì trật tự, an toàn hàng hải. Trong bối cảnh hoạt động hàng hải ngày càng gia tăng cả về quy mô và mức độ phức tạp, các hệ thống giám sát truyền thống dựa trên một loại cảm biến đơn lẻ ngày càng bộc lộ nhiều hạn chế, đặc biệt trong bài toán phát hiện, nhận dạng và phân loại các loại tàu biển. Trước yêu cầu thực tiễn đó, luận án đã tập trung nghiên cứu và đề xuất các phương pháp tiếp cận mới cho bài toán giám sát bờ biển dựa trên sự kết hợp giữa tín hiệu radar xung và dữ liệu ảnh camera, với sự hỗ trợ của các kỹ thuật trí tuệ nhân tạo hiện đại.

Luận án đã tiến hành khảo sát có hệ thống các nghiên cứu trong và ngoài nước liên quan đến giám sát bờ biển, nhận dạng và phân loại tàu biển từ tín hiệu radar cũng như từ dữ liệu ảnh. Kết quả tổng quan cho thấy, mặc dù đã có nhiều công trình đạt được những thành tựu quan trọng, phần lớn các nghiên cứu vẫn tiếp cận các nguồn dữ liệu radar và camera một cách tách biệt. Việc tích hợp hai loại cảm

biến này, đặc biệt trong bối cảnh dữ liệu thực tế ven bờ với nhiều nhiễu và điều kiện quan sát biến đổi, vẫn còn nhiều khoảng trống nghiên cứu. Đây chính là cơ sở khoa học và thực tiễn để luận án đề xuất các hướng tiếp cận mới.

Về phía dữ liệu ảnh, luận án đã đề xuất hai mô hình phát hiện tàu biển dựa trên học sâu. mô hình thứ nhất xây dựng trên nền tảng mạng nơ-ron tích chập với các khối lựa chọn đặc trưng được thiết kế nhằm loại bỏ thông tin dư thừa và tăng cường các đặc trưng mang tính phân biệt cao. Các kết quả thực nghiệm trên các bộ dữ liệu lớn, được công bố rộng rãi cho thấy mô hình đề xuất không chỉ đạt độ chính xác cao hơn so với nhiều phương pháp tiên tiến mà còn đặc biệt hiệu quả trong điều kiện dữ liệu huấn luyện hạn chế. mô hình thứ hai dựa trên kiến trúc transformer, khai thác cơ chế chú ý để tập trung vào các vùng thông tin quan trọng trong ảnh. Nhờ khả năng biểu diễn toàn cục và linh hoạt, mô hình này cho thấy hiệu quả vượt trội trong các kịch bản dữ liệu nhỏ, đồng thời cung cấp các bản đồ chú ý có ý nghĩa trực quan, hỗ trợ phân tích và giải thích kết quả.

Đối với dữ liệu radar xung, luận án đã xây dựng một bộ dữ liệu thực tế thu thập từ trạm radar giám sát ven bờ tại Việt Nam. Đây là một đóng góp quan trọng, bởi phần lớn các nghiên cứu trước đây trong lĩnh vực nhận dạng mục tiêu radar thường dựa trên dữ liệu mô phỏng hoặc dữ liệu hạn chế về mặt kịch bản. Trên cơ sở bộ dữ liệu này, luận án đã kết hợp tri thức chuyên gia với các kỹ thuật học máy và học sâu để rút trích các đặc trưng phản ánh cả yếu tố vật lý của tín hiệu radar và đặc điểm hình dạng sóng liên quan đến loại tàu. Phương pháp phân nhóm và phân loại tàu biển dựa trên tín hiệu radar được đề xuất cho thấy khả năng phân biệt các nhóm tàu chính trong khu vực khảo sát, đồng thời làm rõ những hạn chế của các phương pháp phân nhóm truyền thống khi chúng chủ yếu bị chi phối bởi khoảng cách quan sát thay vì bản chất mục tiêu.

Một đóng góp nổi bật khác của luận án là việc phân tích sâu các thách thức trong bài toán phân nhóm và phân loại dữ liệu radar ven bờ, từ các phương pháp truyền thống đến các phương pháp dựa trên học sâu. Thông qua các phân tích và thực nghiệm, luận án chỉ ra rằng các phương pháp học sâu, khi được thiết kế phù hợp và có sự hướng dẫn của tri thức chuyên gia, có tiềm năng vượt trội trong việc học các biểu diễn đặc trưng mang tính phân biệt cao, ngay cả trong điều kiện dữ liệu nhiễu và không đồng nhất. Những kết quả này không chỉ có ý nghĩa về mặt học thuật mà còn có giá trị tham khảo quan trọng cho việc triển khai các hệ thống giám sát thực tế.

Tổng hợp các kết quả đạt được, luận án đã đóng góp cả về phương diện lý thuyết và thực tiễn cho lĩnh vực giám sát bờ biển. Về mặt lý thuyết, luận án làm rõ vai trò của việc kết hợp đa cảm biến và học sâu trong bài toán nhận dạng và phân loại tàu biển. Về mặt thực tiễn, các mô hình và bộ dữ liệu được xây

dựng trong luận án có thể làm nền tảng cho việc phát triển các hệ thống giám sát bờ biển thông minh, góp phần giảm sự phụ thuộc vào yếu tố con người, nâng cao tính nhất quán và độ tin cậy của quá trình ra quyết định.

Mặc dù đã đạt được những kết quả quan trọng trong việc ứng dụng trí tuệ nhân tạo vào hệ thống cảnh giới bờ biển, luận án vẫn còn bốn hạn chế cốt lõi cần được tiếp tục khắc phục. Thứ nhất, bộ dữ liệu tín hiệu radar xung thực tế được thu thập từ trạm quan trắc tại Việt Nam mới chỉ phân nhóm được ba phân lớp cơ bản gồm tàu cá, tàu vận tải dân dụng và tàu vận tải quân sự, chưa bao quát toàn diện sự đa dạng của các phương tiện hàng hải đang hoạt động trên biển. Thứ hai, toàn bộ quá trình thử nghiệm và đánh giá mô hình hiện tại mới chỉ được thực hiện trên nguồn dữ liệu ngoại tuyến được lưu trữ sẵn dưới dạng các tệp trích xuất thủ công, do đó chưa giải quyết triệt để các thách thức về độ trễ và khả năng xử lý luồng tín hiệu liên tục khi triển khai theo thời gian thực tại các đài radar. Thứ ba, mặc dù luận án đã xây dựng thành công các mô hình nhận dạng độc lập cho tín hiệu radar và dữ liệu hình ảnh quang học, nghiên cứu vẫn chưa thiết lập được một cơ chế hợp nhất đa cảm biến một cách hoàn chỉnh để đưa ra quyết định phân loại cuối cùng trong cùng một kịch bản giám sát. Thứ tư, hiệu suất của các mô hình vẫn chịu ảnh hưởng lớn từ nhiễu môi trường, sự suy hao tín hiệu ở cự ly xa, đồng thời đòi hỏi khối lượng tính toán lớn, tạo ra rào cản về chi phí phần cứng khi triển khai thực tế trên các thiết bị biên. Từ những giới hạn này, định hướng nghiên cứu trong thời gian tới sẽ tập trung vào việc thu thập đa dạng các phân lớp tàu, tối ưu hóa thuật toán xử lý trực tuyến cũng như nén mô hình, và nghiên cứu các kiến trúc học sâu có khả năng dung hợp đồng thời đặc trưng hình thái từ camera và đặc trưng phản xạ từ radar để

Mặt khác, hai hướng tiếp cận độc lập trong luận án đều bộc lộ những ưu, nhược điểm mang tính bổ trợ cho nhau. Phương pháp nhận dạng qua ảnh camera có thế mạnh vượt trội về độ chi tiết và trực quan khi phân loại, nhưng dễ bị suy giảm hiệu năng trong điều kiện thời tiết xấu hay sương mù. Trái lại, phương pháp dùng tín hiệu radar lại hoạt động vô cùng ổn định bất chấp mọi rào cản thời tiết và ánh sáng, song lại thiếu đi các đặc trưng hình học sắc nét để định danh mục tiêu chi tiết. Do đó, hướng nghiên cứu đột phá và thiết thực nhất trong tương lai là hợp nhất đa nguồn dữ liệu (image-radar fusion). Sự giao thoa này sẽ tận dụng tối đa thế mạnh của từng loại cảm biến để bù lấp khiếm khuyết cho nhau, từ đó kiến tạo một hệ thống cảnh giới bờ biển toàn diện, hoạt động bền bỉ và chính xác trong mọi điều kiện môi trường. Nhìn chung, luận án đã giải quyết một bài toán có ý nghĩa khoa học và thực tiễn cao, đáp ứng nhu cầu cấp thiết trong giám sát bờ biển hiện đại. Những kết quả đạt được không chỉ góp phần nâng cao hiệu quả nhận dạng và phân loại tàu biển mà còn mở ra các hướng nghiên cứu

và ứng dụng mới trong phát triển các hệ thống cảnh giới bờ biển thông minh, phục vụ mục tiêu bảo đảm an ninh, an toàn và phát triển bền vững các hoạt động trên biển.

5.2 Kiến nghị

Từ các kết quả nghiên cứu đạt được và những hạn chế còn tồn tại, luận án đề xuất một số kiến nghị nhằm định hướng phát triển và hoàn thiện hơn các hệ thống giám sát bờ biển trong tương lai, đặc biệt theo hướng tích hợp và khai thác hiệu quả dữ liệu từ nhiều loại cảm biến khác nhau.

Thứ nhất, cần tiếp tục nghiên cứu sâu hơn các phương pháp kết hợp dữ liệu đa cảm biến giữa ảnh quang học và radar xung trong bài toán phát hiện, nhận dạng và phân loại tàu biển. Mỗi loại cảm biến đều có những ưu điểm và hạn chế riêng: dữ liệu ảnh cung cấp thông tin trực quan phong phú về hình dạng, kích thước và cấu trúc của tàu, trong khi radar xung có khả năng hoạt động ổn định trong mọi điều kiện thời tiết, ánh sáng và có độ tin cậy cao về mặt phát hiện mục tiêu. Việc kết hợp hai nguồn thông tin này có thể giúp hệ thống giám sát khắc phục được các điểm yếu của từng cảm biến đơn lẻ, từ đó nâng cao độ chính xác và tính ổn định trong các tình huống phức tạp.

"Thứ hai, kiến nghị phát triển các mô hình học sâu hợp nhất đa cảm biến theo nhiều mức độ khác nhau, bao gồm mức dữ liệu (early fusion), mức đặc trưng (late fusion) và mức quyết định (decision-level fusion). Ở mức dữ liệu, cần nghiên cứu các phương pháp đồng bộ không gian – thời gian giữa tín hiệu radar và ảnh camera để tạo ra các cặp dữ liệu có tính tương ứng cao. Ở mức đặc trưng, các kiến trúc mạng đa nhánh có thể được thiết kế để học các biểu diễn đặc trưng riêng cho từng loại cảm biến trước khi hợp nhất, nhằm tận dụng tối đa thông tin bổ sung lẫn nhau. Ở mức quyết định, các chiến lược kết hợp kết quả phát hiện và phân loại từ từng cảm biến có thể giúp hệ thống linh hoạt hơn trong điều kiện một trong hai nguồn dữ liệu bị suy giảm chất lượng. Từ sự phân tích các chiến lược này, hướng phát triển trọng tâm là đề xuất và xây dựng một kiến trúc hệ thống nhận dạng hoàn chỉnh (image-radar fusion architecture), tích hợp xuyên suốt cả hai nguồn dữ liệu từ khâu thu nhận đến quyết định cuối cùng, nhằm tạo ra một giải pháp cảnh giới bờ biển toàn diện, hoạt động bền bỉ và có độ tin cậy cao trong mọi điều kiện môi trường."

Thứ ba, cần mở rộng công tác thu thập và xây dựng các bộ dữ liệu đa cảm biến thực tế trong điều kiện giám sát ven bờ tại Việt Nam. Các bộ dữ liệu này nên bao phủ nhiều kịch bản hoạt động khác nhau như thời tiết xấu, biển động, mật độ tàu cao hoặc mục tiêu nhỏ khó phát hiện. Việc có được các bộ dữ liệu đồng bộ giữa radar và camera không chỉ phục vụ huấn luyện và đánh giá mô hình học sâu mà còn

có ý nghĩa quan trọng trong việc chuẩn hóa và so sánh các phương pháp nghiên cứu trong tương lai.

Thứ tư, kiến nghị tăng cường khai thác tri thức chuyên gia trong quá trình kết hợp hai loại cảm biến. Các hiểu biết về đặc tính phản xạ radar của từng loại tàu, cũng như mối quan hệ giữa hình dạng quan sát được trên ảnh và tín hiệu radar thu được, có thể được sử dụng để thiết kế các ràng buộc hoặc hàm mục tiêu phù hợp cho mô hình học sâu. Điều này giúp nâng cao tính giải thích và độ tin cậy của hệ thống, đặc biệt trong các ứng dụng có yêu cầu cao về an ninh và an toàn.

Thứ năm, Để hướng tới việc ứng dụng vào các hệ thống cảnh giới bờ biển thực tế, định hướng phát triển tiếp theo cần tập trung giải quyết bài toán tối ưu hóa tài nguyên phần cứng. Việc áp dụng các kỹ thuật nén và tăng tốc mô hình là cần thiết để đưa thuật toán lên các hệ thống nhúng với chi phí thấp mà vẫn đảm bảo năng lực xử lý thời gian thực. Đồng thời, nghiên cứu cần mở rộng sang hướng dung hợp dữ liệu (data fusion) đa cảm biến giữa camera và radar. Cơ chế này sẽ giúp hệ thống tự động bù trừ nhược điểm của từng loại thiết bị trước các điều kiện môi trường nhiễu động, qua đó nâng cao độ tin cậy và khả năng giám sát toàn diện 24/7

Tài liệu tham khảo

[1] Q. Chen, Y. Wang, T. Yang, X. Zhang, J. Cheng, and J. Sun, "You only look one-level feature," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 034–13 043.

[2] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," 2021. [Online]. Available: https://arxiv.org/abs/2107.08430

[3] S. Cope *et al.*, "Coastal radar as a tool for continuous and fine-scale monitoring of vessel activities," *PLOS ONE*, 2022, open-access journal article.

[4] "On a new atr tool for coastal surveillance: Pulse-doppler radar applications," conference or technical paper; bibliographic details incomplete.

[5] D. Yang *et al.*, "A review of intelligent ship marine object detection based on rgb camera," *IET Image Processing*, 2024.

[6] "Leveraging deep learning and computer vision for coastal fisheries monitoring," *Scientific Reports*, 2024.

[7] J. Molina *et al.*, "Robust sensor fusion in real maritime surveillance scenarios," 2010, classic paper on maritime sensor fusion.

[8] Y. Zhou *et al.*, "Review on millimeter-wave radar and camera fusion: Techniques and applications," *Sustainability*, 2022.

[9] L. Trinh, S. Mercelis, and A. Anwar, "A comprehensive review of datasets and deep learning techniques for vision in unmanned surface vehicles," *Ocean Engineering*, vol. 334, p. 121501, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801825011850

[10] T. H. Tran, A. Sentchev, T. To Duy, M. Herrmann, S. Ouillon, and K. C. Nguyen, "Surface circulation characterization along the middle southern coastal region of vietnam from high-frequency radar and numerical modeling," *Ocean Science*, vol. 21, no. 1, pp. 1–18, 2025. [Online]. Available: https://os.copernicus.org/articles/21/1/2025/

[11] B. Tran, "Vietnam's quest for enhanced maritime domain awareness," ISEAS – Yusof Ishak Institute, Singapore, ISEAS Perspective 2023/96, December 2023, iSSN 2335-6677. [Online]. Available: https://www.iseas.edu.sg/wp-content/uploads/2024/01/ISEAS_Perspective_2023_96.pdf

[12] N. K. Cuong, T. N. Anh, N. X. Loc, P. D. H. Binh, and V. H. Dang, "Advanced high-resolution measurements of surface waves and currents using two land-based hf radars for offshore operations," in *Proceedings of the 3rd Vietnam Symposium on Advances in Offshore Engineering*, D. V. K. Huynh, H. Doan, T. M. Cao, and P. Watson, Eds. Singapore: Springer Nature Singapore, 2025, pp. 61–69.

[13] N. Mai, A. Sentchev, and T. Cuong, "Applying the method of eof interpolation and 2dvar to complete the ocean surface current data obtained from hf radar system," *VNU Journal of Science: Earth and Environmental Sciences*, vol. 34, 12 2018.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," 2015. [Online]. Available: https://arxiv.org/abs/1506.01497

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision – ECCV 2016*. Springer International Publishing, 2016, pp. 21–37. [Online]. Available: https://doi.org/10.1007%2F978-3-319-46448-0_2

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

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[18] ——, *End-to-End Object Detection with Transformers*, 11 2020, pp. 213–229.

[19] H. W. Kuhn, "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1–2, pp. 83–97, March 1955.

[20] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 658–666.

[21] P. Hinz, "The layer-wise l1 loss landscape of neural nets is more complex around local minima," 2021. [Online]. Available: https://arxiv.org/abs/2105.02831

[22] H. S. Emadi and S. M. Mazinani, "A novel anomaly detection algorithm using dbscan and svm in wireless sensor networks," *Wireless Personal Communications*, vol. 98, pp. 2025–2035, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:39736486

[23] Y. Meng, Y. Zhang, J. Huang, Y. Zhang, C. Zhang, and J. Han, "Hierarchical topic mining via joint spherical tree and text embedding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, ser. KDD '20. ACM, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1145/3394486.3403242

[24] S. Wibisono, M. Anwar, A. Supriyanto, and I. Amin, "Multivariate weather anomaly detection using dbscan clustering algorithm," *Journal of Physics: Conference Series*, vol. 1869, p. 012077, 04 2021.

[25] J. S. Walker, *Fast Fourier transforms*. Boca Raton, Fla.: CRC Press, 1996. [Online]. Available: http://www.amazon.co.uk/gp/search?index=books&linkCode=qs&keywords=9780849371639

[26] N. Ahmed, T. Natarajan, and K. Rao, "Discrete cosine transform," *IEEE Transactions on Computers*, vol. C-23, no. 1, pp. 90–93, 1974.

[27] M. Fu, H. Liu, Y. Yu, J. Chen, and K. Wang, "Dw-gan: A discrete wavelet transform gan for nonhomogeneous dehazing," 2021.

[28] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol. 74, no. 3, pp. 646–648, 1987. [Online]. Available: http://biomet.oxfordjournals.org/content/74/3/646.abstract

[29] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[30] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, "Euclidean distance geometry and applications," 2012.

[31] S. R. Blackburn, C. Homberger, and P. Winkler, "The minimum manhattan distance and minimum jump of permutations," 2018.

[32] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *JSTOR: Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[33] P. Macgregor, "Fast and simple spectral clustering in theory and practice," 2023.

[34] D.-y. Xia, F. Wu, X.-q. Zhang, and Y.-t. Zhuang, "Local and global approaches of affinity propagation clustering for large scale data," *Journal of Zhejiang University-SCIENCE A*, vol. 9, no. 10, p. 1373–1381, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1631/jzus.A0720058

[35] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2021.

[36] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, p. 307–392, 2019. [Online]. Available: http://dx.doi.org/10.1561/2200000056

[37] X. Liu, F. Zhang, Z. Hou, Z. Wang, L. Mian, J. Zhang, and J. Tang, "Self-supervised Learning: Generative or Contrastive," *arXiv:2006.08218 [cs, stat]*, Jul. 2020, arXiv: 2006.08218. [Online]. Available: http://arxiv.org/abs/2006.08218

[38] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features." *CoRR*, vol. abs/1807.05520, 2018. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1807.html#abs-1807-05520

[39] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015.

[40] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," 2016.

[41] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," 2014.

[42] J. Zhao, M. Mathieu, R. Goroshin, and Y. LeCun, "Stacked what-where auto-encoders," 2016.

[43] S. J. Wetzel, "Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders," *Physical Review E*, vol. 96, no. 2, Aug. 2017. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.96.022140

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[45] J. Zheng and Y. Liu, "A study on small-scale ship detection based on attention mechanism," *IEEE Access*, vol. 10, pp. 77 940–77 949, 2022.

[46] B. Ye, T. Qin, H. Zhou, J. Lai, and X. Xie, "Cross-level attention and ratio consistency network for ship detection," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022, pp. 4644–4650.

[47] H. Cui, Y. Yang, M. Liu, T. Shi, and Q. Qi, "Ship detection: An improved yolov3 method," in *OCEANS 2019 - Marseille*, 2019, pp. 1–4.

[48] T. Liu, B. Pang, S. Ai, and X. Sun, "Study on visual detection algorithm of sea surface targets based on improved yolov3," *Sensors*, vol. 20, no. 24, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/24/7263

[49] H. Li, L. Deng, C. Yang, J. Liu, and Z. Gu, "Enhanced yolo v3 tiny network for real-time ship detection from visual image," *IEEE Access*, vol. 9, pp. 16 692–16 706, 2021.

[50] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016. [Online]. Available: https://arxiv.org/abs/1612.03144

[51] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[52] S. Woo, J. Park, J.-Y. Lee, and I. Kweon, "Cbam: Convolutional block attention module: 15th european conference, munich, germany, september 8–14, 2018, proceedings, part vii," 09 2018, pp. 3–19.

[53] T. Liu, B. Pang, L. Zhang, W. Yang, and X. Sun, "Sea surface object detection algorithm based on yolo v4 fused with reverse depthwise separable convolution (rdsc) for usv," *Journal of Marine Science and Engineering*, vol. 9, no. 7, 2021. [Online]. Available: https://www.mdpi.com/2077-1312/9/7/753

[54] J. Guo, Y. Li, W. Lin, Y. Chen, and J. Li, "Network decoupling: From regular to depthwise separable convolutions," 2018.

[55] X. Han, L. Zhao, Y. Ning, and J. Hu, "Shipyolo: An enhanced model for ship detection," *Journal of Advanced Transportation*, vol. 2021, pp. 1–11, 06 2021.

[56] M. Zhang, X. Rong, and X. Yu, "Light-sdnet: A lightweight cnn architecture for ship detection," *IEEE Access*, vol. 10, pp. 86 647–86 662, 2022.

[57] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," 2020.

[58] R. Ye, F. Liu, and L. Zhang, "3d depthwise convolution: Reducing model parameters in 3d vision tasks," 2018.

[59] Q. Zhang, Y. Huang, and R. Song, "A ship detection model based on yolox with lightweight adaptive channel feature fusion and sparse data augmentation," in *2022 18th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2022, pp. 1–8.

[60] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[61] Y. Zhang, M. J. Er, W. Gao, and J. Wu, "High performance ship detection via transformer and feature distillation," in *2022 5th International Conference on Intelligent Autonomous Systems (ICoIAS)*, 2022, pp. 31–36.

[62] Z. Zhang, L. Zhang, Y. Wang, P. Feng, and R. He, "Shiprsimagenet: A large-scale fine-grained dataset for ship detection in high-resolution optical remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 8458–8472, 2021.

[63] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999, pp. 368–377. [Online]. Available: https://arxiv.org/abs/physics/0004057

[64] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948. [Online]. Available: http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

[65] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," 2016. [Online]. Available: http://arxiv.org/abs/1612.00410

[66] "Seaship dataset," 10, 20, 2022. [Online]. Available: https://www.kaggle.com/datasets/tangwenyang/seaship

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[68] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," 2021.

[69] M. Lueken, W. ten Kate, J. P. Batista, C. Ngo, C. Bollheimer, and S. Leonhardt, "Peak detection algorithm for gait segmentation in long-term monitoring for stride time estimation using inertial measurement sensors," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, 2019, pp. 1–4.

[70] L. Du, L. Li, B. Wang, and J. Xiao, "Micro-doppler feature extraction based on time-frequency spectrogram for ground moving targets classification with low-resolution radar," *IEEE Sensors Journal*, vol. 16, pp. 1–1, 05 2016.

[71] M. E. Demirhan and O. Salor, "Classification of targets in sar images using svm and k-nn techniques," in *2016 24th Signal Processing and Communication Application Conference (SIU)*, 2016, pp. 1581–1584.

[72] M. Ren, J. Cai, Y. Zhu, and M. He, "Radar emitter signal classification based on mutual information and fuzzy support vector machines," in *2008 9th International Conference on Signal Processing*, 2008, pp. 1641–1646.

[73] C. Wang, J. Pei, R. Wang, Y. Huang, and J. Yang, "A new ship detection and classification method of spaceborne sar images under complex scene," in *2019 6th Asia-Pacific Conference on Synthetic Aperture Radar (APSAR)*, 2019, pp. 1–4.

[74] U. Kaydok, "Chaff discrimination using convolutional neural networks and range profile data," in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 373–377.

[75] S. Rajkamal, "Selecting reviewers for research by clustering proposals using expectation maximization clustering algorithm," in *2017 International Conference on Technical Advancements in Computers and Communications (ICTACC)*, 2017, pp. 56–60.

[76] A. E. Vincent and K. Sreekumar, “A survey on approaches for ecg signal analysis with focus to feature extraction and classification,” in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2017, pp. 140–144.

[77] S.-S. Kim and T. Kasparis, “A modified domain deformation theory on 1-d signal classification,” *IEEE Signal Processing Letters*, vol. 5, no. 5, pp. 118–120, 1998.

[78] M. S. Azmi, N. A. Arbain, A. K. Muda, Z. A. Abas, and Z. Muslim, “Data normalization for triangle features by adapting triangle nature for better classification,” in *2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2015, pp. 1–6.

DANH MỤC CÔNG TRÌNH CỦA NGHIÊN CỨU SINH LIÊN QUAN ĐẾN ĐỀ TÀI

- **Duc-Dat Ngo**, Van-Linh Vo, Tri Nguyen, Manh-Hung Nguyen, & My-Ha . (2023). Image-Based Ship Detection Using Deep Variational Information Bottleneck. Sensors, 23(19), 8093. https://doi.org/10.3390/s23198093. Q1; ISSN: 1424-8220
- **Duc-Dat Ngo**, Van-Linh Vo, My-Ha Le, Hoc-Phan, & Manh-Hung Nguyen (2024). Transformer based ship detector: An improvement on feature map and tiny training set. EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, 12(1). https://doi.org/10.4108/eetinis.v12i1.6794. Q3; ISSN: 2410-0218
- **Duc-Dat Ngo**, Manh-Hung Nguyen, Quang-Thai-Dan Nguyen., & My-Ha Le. (2021). Clustering based ship classification using radar signal and neuron network. In Proceedings of the 2021 International Conference on System Science and Engineering (ICSSE) (pp. 122-127). IEEE. https://doi.org/10.1109/ICSSE52999.2021.9538475. ISSN: 2325-0925 ISBN:978-1-6654-4848-2
- **Duc-Dat Ngo**, Van-Hoang-Anh Phan,Huynh-The Pham, Tien-Tan Be, Van-Binh Nguyen, & My-Ha Le (2023). A vision-based container-code checking system: Case study at international terminal. In Proceedings of the 2023 International Workshop on Intelligent Systems (IWIS) (pp. 1–6). IEEE. https://doi.org/10.1109/IWIS58789.2023.10284525. ISBN:979-8-3503-0504-3