

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

CAO LÊ MINH HIẾU

PHÁT HIỆN ĐỐI TƯỢNG TRỌNG YẾU

DỰA VÀO MẠNG NƠ RON TÍCH CHẬP

*(Salient Object Detection Based on Deep Convolutional
Neural Network)*

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

CHƯƠNG TRÌNH CHÍNH QUY

Tp. Hồ Chí Minh, tháng 07/2022

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

CAO LÊ MINH HIẾU – 18120368

**PHÁT HIỆN ĐỐI TƯỢNG TRỌNG YẾU
DỰA VÀO MẠNG NƠ RON TÍCH CHẬP**

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN CNTT

CHƯƠNG TRÌNH CHÍNH QUY

GIÁO VIÊN HƯỚNG DẪN

TS. Nguyễn Đức Hoàng Hạ

Tp. Hồ Chí Minh, tháng 07/2022

Lời cam đoan

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của TS. Nguyễn Đức Hoàng Hạ. Các số liệu và kết quả nghiên cứu trong luận văn này là trung thực và không có bất cứ sự sao chép, trùng lặp với các đề tài khác.

Lời cảm ơn

Em xin trân trọng cảm ơn các quý thầy cô trong khoa Công Nghệ Thông Tin trường Đại học Khoa Học Tự Nhiên Tp. Hồ Chí Minh đã tận tâm giảng dạy, truyền đạt cả kiến thức lẫn kinh nghiệm trong các năm học của em trên ghế nhà trường và tạo điều kiện cho em thực hiện khóa luận tốt nghiệp.

Dưới sự hướng dẫn và chỉ dạy nhiệt tình của thầy Nguyễn Đức Hoàng Hạ, khóa luận của em đã có thể hoàn thành. Em xin chân thành cảm ơn thầy.

Đồng thời em cũng xin cảm ơn sự giúp đỡ nhiệt tình của thầy Lý Quốc Ngọc đã dẫn dắt em trong những lần em đi sai hướng. Một lần nữa, em xin cảm ơn hai thầy rất nhiều.

Cảm ơn cha mẹ, người thân và bạn bè đã hết lòng động viên, chia sẻ những khó khăn trong quá trình thực hiện khóa luận tốt nghiệp.

Em đã cố gắng hoàn thành khóa luận tốt nghiệp tốt nhất có thể trong phạm vi kiến thức và khả năng của mình nên chắc chắn sẽ không tránh được những thiếu sót trong lần đầu nghiên cứu. Em rất mong các quý thầy cô thông cảm và dạy bảo nhiệt tình.

Đề cương chi tiết

ĐỀ CƯƠNG KHÓA LUẬN TỐT NGHIỆP PHÁT HIỆN ĐỐI TƯỢNG TRỌNG YẾU DỰA VÀO MẠNG NƠ RON TÍCH CHẬP

*(Salient Object Detection Based on Deep
Convolutional Neural Network)*

THÔNG TIN CHUNG

- **Người hướng dẫn:**
 - TS Nguyễn Đức Hoàng Hạ (Khoa Công Nghệ Thông Tin, Bộ môn Thị giác máy tính và Điều khiển học thông minh)
- **Sinh viên thực hiện:**
 - Cao Lê Minh Hiếu (MSSV:18120368).

Loại đề tài: Nghiên cứu.

Thời gian thực hiện: Từ 01/2022 đến 06/2022.

NỘI DUNG THỰC HIỆN

Giới thiệu về đề tài

Phát hiện đối tượng trọng yếu (Salient Object Detection, viết tắt là SOD) là bài toán phát hiện những đối tượng thu hút sự chú ý của con người (Visual Attention), lĩnh vực đã được nghiên cứu rất lâu bởi các nhà nghiên cứu về khoa học nhận thức. Các

nhà khoa học thị giác máy tính mong muốn huấn luyện được mô hình có thể mô phỏng hệ thống thị giác của con người để đánh giá mức độ quan trọng của đối tượng trong bức ảnh. SOD được xem như là bước tiền xử lý trong rất nhiều tác vụ của thị giác máy tính như image retrieval, visual tracking, image segmentation, person re-identification, content-aware image editing.

Vì có nguồn gốc từ khoa học nhận thức nên việc hiểu rõ cơ chế Visual Attention vẫn là một câu hỏi lớn. Với những bức ảnh chỉ có một đối tượng, hoặc đối tượng đó chiếm vị trí hầu như cả bức ảnh thì con người dễ dàng nhận biết chúng là trọng yếu. Nhưng một bức ảnh, khung hình bao gồm nhiều đối tượng thì việc nhìn nhận một đối tượng là trọng yếu sẽ khác nhau giữa người với người.

Ví dụ như những người có xu hướng nhìn tổng quan trước, chi tiết sau sẽ xem các vật thể lớn là trọng yếu, và ngược lại. Hay việc người đứng ở giữa thu hút sự chú ý nhất chiếm tỉ lệ cao nhưng không phải là hoàn toàn, sẽ có những người lại nhìn về một phía trước khi nhìn vào trung tâm.

Do bài toán mang tính chất chủ quan như vậy, việc tạo ra dữ liệu để huấn luyện và đánh giá mô hình cũng rất khó mang tính nhất quán.

Để mang tính tổng quát nhất có thể, các bộ dữ liệu chuẩn được tạo ra bằng cách như sau:

1. Chọn n người để đánh nhãn, những người này có thị lực, tâm lý bình thường.
2. Với mỗi bức ảnh, mỗi người sẽ tự chọn ra k (bị giới hạn) vật thể được xem là trọng yếu
3. Tổng hợp sự lựa chọn của n người để chọn ra các vật thể được xem là trọng yếu nhiều nhất.
4. Đánh nhãn cho ảnh.

Phương pháp tạo ra bộ dữ liệu mô phỏng Visual Attention của con người tốt nhất có thể là cho chính con người nhận biết chúng. Việc lọc ra các vật thể được đa số người công nhận là trọng yếu đã giảm bớt tính chủ quan của bài toán nhưng thử

thách vẫn còn đó (**hình 1**). Hơn thế nữa, việc huấn luyện và kiểm nghiệm trên các bộ dữ liệu của các mô hình cũng cho kết quả đồng nhất.



(a)



(b)

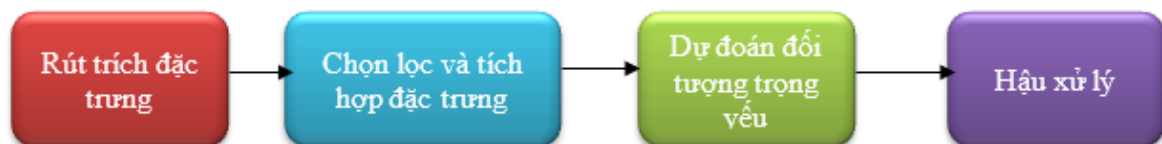
Hình 1: Những ví dụ về việc đánh nhãn thiếu sự nhất quán. **(a)**Tất cả em bé trong ảnh bên trái đều được đánh nhãn, nhưng những người trong ảnh tay phải thì không. **(b)**Ảnh bên tay trái, cả con bướm và bông hoa đều được xem là trọng yếu nhưng ảnh bên tay phải thì không.

Để có bộ dữ liệu tốt hơn nữa yêu cầu vừa nguồn lực, tiền bạc lẫn các nghiên cứu xa hơn về khoa học nhận thức.

Từ những ngày đầu, các phương pháp truyền thống, thủ công được áp dụng để rút trích đặc trưng nhưng chỉ mang tính chất chủ quan, trực giác đã không đạt được độ chính xác như mong đợi. Với sự ra đời của mạng học sâu convolutional neural networks, bức ảnh qua các kiến trúc VGG, Resnet đã có thể được rút trích thông tin cục bộ ở các lớp convolutions nông, thông tin toàn cục khi càng đi sâu tới cuối kiến trúc. Các mô hình hiện đại nhất của từng tác vụ thị giác máy tính đã áp dụng các kiến trúc trên như một hệ thống trích xuất đặc trưng (hay còn được gọi là backbone của mô hình) và đạt những kết quả tốt ngoài mong đợi.

Mô hình phát hiện đối tượng trọng yếu sử dụng mạng học sâu nói chung gồm các quá trình rút trích đặc trưng, chọn lọc và tích hợp đặc trưng, dự đoán đối tượng trọng yếu, hậu xử lý (**hình 2**).

Backbone của các mô hình SOD sử dụng VGG, Resnet để rút trích đặc trưng. Sau đó sẽ chọn lọc bằng các cơ chế attention và tích hợp các đặc trưng sao cho có độ phân giải bằng với ảnh đầu vào. Áp dụng 1x1 convolution theo sau là hàm 2-way softmax lên khối đặc trưng sau khi chọn lọc để dự đoán các vật thể trọng yếu. Giai đoạn hậu xử lý thông thường sẽ nhằm tinh chỉnh biên cạnh vật thể bằng cách phương pháp như conditional random field hoặc một mạng CNNs. Các mô hình tốt nhất hiện nay đã lược bỏ bước hậu xử lý để tăng hiệu quả mà vẫn độ chính xác không bị giảm đi.



Hình 2: SOD pipeline.

Hiện tại mô hình tốt nhất hiện nay cho bài toán phát hiện đối tượng trọng yếu trong ảnh màu hai chiều là TRACER (dựa trên bảng xếp hạng của trang paperswithcode.com) có thể phân vùng những vật thể trọng yếu vô cùng tốt đến cả từng biên cạnh của vật thể. Do đó để có thể dự đoán chính xác đến từng pixel thì việc rút trích và kết hợp các khối đặc trưng (feature maps) từ các khối convolution (convolution blocks) của các backbones một cách có chọn lọc và có độ phân giải bằng với ảnh đầu vào là vô cùng cần thiết. Các bài toán pixelwise dense prediction (phân đoạn ảnh, phát hiện đối tượng trọng yếu) đã nghiên cứu rất nhiều cơ chế tích hợp các đặc trưng bậc cao, độ phân giải thấp với các đặc trưng bậc thấp, độ phân giải cao để lọc ra đặc trưng giàu thông tin liên quan nhất đến bài toán cho từng pixel.

Mặc dù là một nhánh nghiên cứu của object detection – một tác vụ đã được nghiên cứu, phát triển và đạt được những kết quả vô cùng tốt nhưng giữa chúng vẫn có sự khác nhau (**Bảng 1**) dẫn đến việc áp dụng các kỹ thuật, phương pháp từ object detection sang SOD và ngược lại là một thách thức. Chẳng hạn như sau khi phát hiện được các vật thể nhưng việc xác định đối tượng nào trọng yếu là vô cùng khó khăn.

Mục tiêu đề tài

Đề tài nhằm khảo sát các phương pháp học sâu để rút trích, chọn lọc và tích hợp các đặc trưng giàu thông tin cho từng pixel (vừa cục bộ vừa toàn cục để có thể xác định pixel đó thuộc vật thể trọng yếu hay không).

Đề tài cũng sẽ thực nghiệm cơ chế self-attention nhằm khảo sát sự hiệu quả trong việc cung cấp thêm thông tin cho từng pixel về độ quan trọng của chúng trong bức ảnh. Từ hai thực nghiệm trên có thể đề xuất mô hình đạt độ chính xác trên từng pixel.

Phạm vi của đề tài

Phát hiện đối tượng trọng yếu chia thành các bài toán nhỏ như 2D RGB Salient Object Detection, 3D RGB SOD, 4D RGB SOD. Đề tài chỉ tập trung vào bài toán phát hiện đối tượng trọng yếu trên ảnh màu hai chiều 2D RGB SOD.




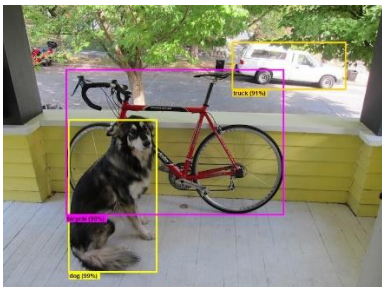
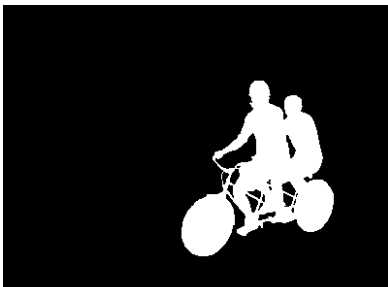

Các mô hình được khảo sát là các mô hình có sử dụng mạng học sâu và phương thức học có giám sát.

Đề tài chỉ tập trung vào các kỹ thuật saliency map, attention map, superpixel được áp dụng vào tác vụ phát hiện đối tượng hiện nay để cải thiện độ chính xác. Phân tích, so sánh giữa các kỹ thuật trên với nhau và nếu có thể cải thiện một mô hình object detection nào đó dựa trên các phương pháp vừa đề cập.

Cách tiếp cận dự kiến

Nghiên cứu sẽ được tiếp cận theo các bước sau:

- Nắm bắt và hiểu được khó khăn của bài toán trong việc định nghĩa một đối tượng quan trọng.
- Tìm hiểu các công trình liên quan, các tập dữ liệu mới thách thức hơn được xây dựng như thế nào.
- Thực nghiệm mô hình đề xuất là kết hợp giữa PFSNet và OCR.
- Áp dụng các hàm lỗi đánh mạnh trọng số vào biên cạnh của vật thể trọng yếu.
- So sánh kết quả của mô hình được đề xuất với những mô hình hiện nay trên các tập dữ liệu chuẩn.

	Phát hiện đối tượng	Phát hiện đối tượng trọng yếu	Phân đoạn/vùng ảnh
Đầu vào	<p>Ảnh màu. Ví dụ:</p> 	<p>Ảnh màu có chứa vật thể trọng yếu. Ví dụ:</p> 	<p>Ảnh màu. Ví dụ:</p> 
Đầu ra	<p>Tọa độ của bounding boxes và class scores cho từng bounding box.</p> 	<p>Một binary mask kích thước bằng ảnh đầu mà giá trị của từng pixel được gán nhãn thuộc các vật thể trọng yếu (không phân biệt sự khác nhau giữa các vật thể) hoặc background.</p> 	<p>Một ma trận mask mà giá trị của từng pixel được gán nhãn tương ứng với từng lớp vật thể.</p> 

Bảng 1: Sự khác nhau giữa các tác vụ phát hiện đối tượng, phát hiện đối tượng trọng yếu và phân đoạn/vùng ảnh.

Kết quả dự kiến của đề tài

Với mục tiêu và cách tiếp cận được nêu ra ở trên, nhóm nghiên cứu mong muốn đạt được các kết quả sau:

- Xây dựng được mô hình đề xuất, hi vọng có thể dự đoán chính xác trên từng điểm ảnh.
- Cho thấy sự hiệu quả của phương pháp vốn được áp dụng trong phân đoạn ảnh có thể cải thiện độ chính xác của mô hình. Từ đó có thể xem xét đưa thêm các kỹ thuật của phân đoạn ảnh vào phát hiện đối tượng trọng yếu.
- Khảo sát kết quả của mô hình với các mô hình hiện nay.

Kế hoạch thực hiện

- **01/01/2022 đến 15/03/2022:** Tìm hiểu các công trình liên quan.
- **16/03/2022 đến 15/05/2022:** Cài đặt mô hình.
- **16/05/2022 đến 22/06/2022:** So sánh kết quả mô hình và hoàn chỉnh cuốn luận.

Danh mục các kí hiệu, chữ viết tắt

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT ĐƯỢC SỬ DỤNG TRONG KHÓA LUẬN

Ký hiệu	Ý nghĩa
SOD	Salient Object Detection
CRF	Conditional Random Field
CNN	Convolutional Neural Network
MAE	Mean Absolute Error
PFSNet	Pyramidal Feature Shrinking Network
OCR	Object-Contextual Representations
MLP	Multi-layer Perceptron
BCE	Binary Cross Entropy
FCN	Fully Convolutional Network
IoU	Intersection over Union
SSIM	Structural Similarity Index Measure
ViT	Vision Transformer

DANH MỤC CÁC TỪ CHUYÊN MÔN ĐƯỢC DỊCH SANG TIẾNG VIỆT

Từ chuyên môn dịch sang tiếng Việt	Từ chuyên môn tiếng Anh
Khoa học nhận thức	Cognitive Science
Phát hiện đối tượng trọng yếu	Salient Object Detection
Phát hiện đối tượng	Object Detection
Mạng tích chập	Convolutional Neural Network
Mạng nơ-ron nhiều lớp	Multi-layer Perceptron
Khối đặc trưng	Feature map
Khối tích chập	Convolutional block
Vùng nhìn thấy của nơ-ron	Receptive Field
Hiện tượng suy giảm đặc trưng khi tích hợp.	Leaping Feature Fusion Operation
Giám sát sâu	Deep Supervision
Kiến trúc đa luồng	Multi-stream architecture
Nội suy song tuyến tính	Bilinear interpolation
Đặc trưng đa bậc	Multi-level feature
Kết nối tắt	Skip connection
Phép cộng với từng phần tử	Element-wise addition
Phép nhân với từng phần tử	Element-wise multiplication
Tích chập giãn nở	Dilated Convolution
Bản đồ điểm quan trọng	Saliency map
Độ chính xác	Precision
Độ phủ	Recall
Giá trị độ lỗi trung bình tuyệt đối	Maximum Absolute Error
Tốc độ học	Learning rate
Kích thước lô	Batch size

Mục lục

Lời cam đoan	i
Lời cảm ơn	ii
Đề cương chi tiết.....	iii
Danh sách các hình.....	xvii
Danh sách các bảng.....	xix
Lời mở đầu.....	xx
Tóm tắt	xxii
Chương 1 Tổng quan	1
1.1 Động lực nghiên cứu	1
1.1.1 Ý nghĩa khoa học	1
1.1.2 Ý nghĩa thực tiễn.....	1
1.2 Phát biểu bài toán	2
1.3 Thách thức của bài toán.....	3
1.3.1 Định nghĩa đối tượng trọng yếu và cách thức đánh nhãn.....	3
1.3.2 Tích hợp đặc trưng.....	8
1.3.3 Áp dụng các phương pháp từ các lĩnh vực tương tự.....	8
1.4 Hướng tiếp cận.....	10
1.5 Đóng góp	10
Chương 2 Các nghiên cứu liên quan	11
2.1 Các nghiên cứu liên quan	11
2.2 Phân tích hướng phát triển và cải tiến	23
2.3 Kết luận.....	24

Chương 3	Phương pháp đề xuất	25
3.1	Rút trích đặc trưng	25
3.2	Chọn lọc và tích hợp đặc trưng.....	28
3.2.1	Scale-aware Enrichment Module (SEM).....	29
3.2.2	Adjacent Fusion Module (AFM)	31
3.2.3	Cơ chế channel-attention (WCAT).....	33
3.2.4	Cơ chế dựa trên self-attention (OCR).....	35
3.3	Tính toán hàm lỗi.....	39
3.4	Kết luận.....	40
Chương 4	Kết quả thử nghiệm.....	41
4.1	Công tác chuẩn bị dữ liệu	41
4.1.1	Bộ dữ liệu cho quá trình huấn luyện.....	41
4.1.2	Các bộ dữ liệu cho quá trình kiểm thử.....	42
4.2	Phương pháp đánh giá	45
4.2.1	Độ chính xác – Độ phủ	45
4.2.2	F-measure.....	45
4.2.3	MAE – Giá trị độ lỗi trung bình tuyệt đối	46
4.2.4	S-measure.....	46
4.2.5	Kết luận.....	46
4.3	Chi tiết cài đặt, quá trình tinh chỉnh và huấn luyện.....	47
4.4	Đánh giá và so sánh các cấu hình.....	48
4.5	Kết luận.....	57
Chương 5	Tổng kết và hướng phát triển.....	58
5.1	Tổng kết.....	58

5.2	Các hướng nghiên cứu trong tương lai	59
Tài liệu tham khảo		61
Phụ lục		64

Danh sách các hình

Hình 1.2.1: Ảnh đầu vào (bên trái) và nhãn (bên phải) của bài toán phát hiện đối tượng trọng yếu trên ảnh màu hai chiều. Hai bức ảnh này được lấy từ tập DUTS-TR cho việc huấn luyện mô hình ở chương 3.	2
Hình 1.3.1: Cách thức đánh nhãn được thực hiện trong các tập dữ liệu chuẩn dòng 1 ASD , dòng 2 SOD , dòng 3 DUT-OMRON , dòng 4 PASCAL-S	6
Hình 1.3.2: Những ví dụ về việc đánh nhãn thiếu sự nhất quán. Tất cả em bé trong ảnh dòng 1 đều được đánh nhãn, nhưng những người trong ảnh dòng 2 thì không. Ảnh dòng 3, cả con bướm và bông hoa đều được xem là trọng yếu nhưng ảnh dòng 4 phải thì không.	7
Hình 3.1.1: Các viên sỏi trong ảnh bên trái có độ tương phản so với các điểm ảnh xung nhưng chúng không được xem là trọng yếu. Để phát hiện vật thể trọng yếu thì yêu cầu phải rút ra được các đặc trưng vừa mang thông tin cục bộ, vừa mang thông tin toàn cục, giàu ngữ nghĩa của bức ảnh.	26
Hình 3.2.1: Quá trình tích hợp đặc trưng trong mô-đun SEM.	31
Hình 3.2.2: Quá trình tích hợp đặc trưng trong mô-đun AFM.	32
Hình 3.2.3: Kiến trúc PFSNet[12] để rút trích và tích hợp các đặc trưng đa bậc. Mô-đun WCAT đóng vai trò như một cơ chế channel attention, giúp chọn lọc đặc trưng. Phép toán với từng phần tử trong mô-đun nhằm nâng cao và tăng cường các đặc trưng tương đồng.	34
Hình 3.2.4: Cơ chế hoạt động Object-Contextual Representations(OCR) [11] dựa trên self-attention	38
Hình 3.3.1: Ảnh đầu vào (bên trái), saliency map (giữa), trọng số (bên phải).	39
Hình 4.1.1: Dữ liệu đánh giá gồm các ảnh có nhiều vật thể trọng yếu.	43
Hình 4.1.2: Vật thể trọng yếu có màu tương đồng với phông nền xung quanh.	44
Hình 4.4.1: Mô hình cải tiến cho loại bỏ các vùng, điểm ảnh không trọng yếu, trong khi giữ lại và tăng cường các đối tượng trọng yếu.	51

Hình 4.4.2: Mô hình cải tiến cho kết quả phân lớp tốt hơn nhiều so với PFSNet trên các bức ảnh bị che hoặc cùng màu với phong nền (occlusion).....	53
Hình 4.4.3: Kết quả trên các bức ảnh chứa vật thể trọng yếu không toàn vẹn, vị trí sát biên ảnh, lớn hơn nửa bức ảnh.....	54
Hình 4.4.4: Kết quả trên các bức ảnh chứa một hoặc nhiều vật thể có kích thước nhỏ.....	55
Hình 4.4.5: Kết quả trên các bức ảnh chứa nhiều vật thể có hình dáng mỏng.	56

Danh sách các bảng

Bảng 1.3.1: Cách thức đánh nhãn cho các bộ dữ liệu chuẩn được sử dụng phổ biến rộng rãi nhất hiện nay.....	5
Bảng 1.3.2: Sự khác nhau giữa các tác vụ phát hiện đối tượng, phát hiện đối tượng trọng yếu và phân đoạn/vùng ảnh.	9
Bảng 2.1.1: Tổng quan các nghiên cứu liên quan.	15
Bảng 2.1.2: Phân tích ưu nhược điểm của các công trình liên quan.	22
Bảng 3.1.1: So sánh kiến trúc Resnet-50 với VGG-16.	27
Bảng 3.1.2: Một phần của kiến trúc Resnet-50 được sử dụng để rút trích đặc trưng của ảnh đầu vào với kích thước HxWx3 (lớp average pool và fully connected được vứt bỏ). Kết quả của quá trình thu được 5 khối đặc trưng như L1, L2, L3, L4, L5 và kích thước đầu ra tương ứng	28
Bảng 3.2.1: Các khối đặc trưng và kích thước của chúng sau khi được tích hợp trong mô-đun SEM.	30
Bảng 3.2.2: Áp dụng OCR từ phân đoạn ngữ nghĩa cho bài toán phát hiện đối tượng trọng yếu.	36
Bảng 4.3.1 Chi tiết quá trình tinh chỉnh và huấn luyện trên các cấu hình khác.....	48
Bảng 4.4.1: Kết quả đánh giá các mô hình trên các bộ dữ liệu kiểm thử. 64, 128 lần lượt là số kênh trong mô-đun OCR, theo hai cấu hình nhóm khảo sát.	49

Lời mở đầu

Trong thời đại mà thị giác máy tính sử dụng mạng học sâu phát triển cực kì nhanh, con người đã xây dựng được các mô hình vô cùng tốt, vượt qua khả năng của con người trong một số tác vụ như phân loại hình ảnh, phát hiện đối tượng. Các mô hình học sâu này hoạt động hiệu quả là nhờ vào sự dồi dào của dữ liệu thị giác đã được đánh nhãn. Những lĩnh vực thử thách và khó khăn hơn dần chiếm lấy sự chú ý của các học giả, kỹ sư và các nhà nghiên cứu. Phát hiện đối tượng trọng yếu là một trong số lĩnh vực thị giác đang được cộng đồng quan tâm trong những năm gần đây và đã đạt được nhiều kết quả vô cùng ấn tượng.

Phát hiện đối tượng trọng yếu được xem là một bài toán khó bởi vì nhận định như thế nào là trọng yếu mang tính chủ quan rất cao. Con người sinh ra không ai giống ai, quan điểm, sở thích và khả năng nhận thức của mỗi người là khác nhau. Do đó, với các vật thể trước tầm nhìn của con người, mỗi cá nhân sẽ bị các vật thể khác nhau thu hút.

Nhận thấy được khó khăn, các nhóm nghiên cứu trên thế giới đã tận dụng khả năng tự học của các mô hình học sâu hiện đại trên các tập dữ liệu chuẩn được đánh nhãn bởi chính con người có khả năng nhận thức bình thường. Từ đó, mô hình thu được sẽ phát hiện vật thể trọng yếu dựa vào những gì đã được huấn luyện trên tập dữ liệu.

Nhóm nghiên cứu lựa chọn các mô hình học sâu bởi thứ nhất chúng cho thấy khả năng rút trích được các đặc trưng bậc cao, phức tạp, giàu ngữ nghĩa của bức ảnh. Khả năng đó không tìm thấy được trong các kỹ thuật học máy phụ thuộc nhiều vào các đặc trưng được rút trích thủ công dựa vào chủ yếu kiến thức lĩnh vực đó. Thứ hai, các bài báo khoa học về đề tài tập trung chủ yếu vào kỹ thuật học sâu và đã đạt các kết quả vô cùng tốt. Từ đó sẽ là nguồn tìm hiểu dồi dào cho nhóm nghiên cứu thực hiện khóa luận cũng như tiếp cận với các kỹ thuật mới nhất hiện nay.

Trong khóa luận này, nhóm sẽ khảo sát các mô hình học sâu cũng như các tập dữ liệu tạo ra các cột mốc độ phát trong các năm. Nhóm hi vọng sau khi nghiên cứu các bài báo liên quan trên sẽ đưa ra được các ý tưởng, cải tiến và từ đó có thể đóng góp vào quá trình phát triển của đề tài.

Tóm tắt

Chương đầu tiên khóa luận đưa ra động lực nghiên cứu, giới thiệu bài toán cũng như những thách thức mà đề tài đang phải giải quyết. Ở chương tiếp theo, nhóm trình bày quá trình phát triển của đề tài qua các năm, từ đó chỉ ra các ưu, nhược điểm của các nghiên cứu liên quan và rút ra được xu hướng nghiên cứu. Nhóm lựa chọn mô hình khảo sát PFSNet chỉ tích hợp các khối đặc trưng liền kề, có sự cách biệt nhỏ nhất về kích thước và mức độ thông tin. Lần đầu tiên áp dụng mô-đun Object Contextual Representation (OCR) trong bài toán phân đoạn ngữ nghĩa thành công cho đề tài. Mô-đun dựa trên cơ chế self-attention nhằm điều chỉnh mức độ trọng yếu của từng điểm ảnh dựa vào mức độ trọng yếu của tất cả các điểm ảnh khác. Chi tiết mô hình gốc PFSNet và mô-đun cải tiến OCR được phân tích ở chương 3. Quá trình chuẩn bị dữ liệu và huấn luyện mô hình cải tiến được đề cập ở chương 4. Các độ đo, bộ dữ liệu được sử dụng để huấn luyện, đánh giá và kết quả thử nghiệm cũng được trình bày trong chương này. Khóa luận kết thúc bằng việc tổng kết quá trình thực hiện, đóng góp của nhóm và đưa ra các hướng phát triển trong tương lai.

Chương 1

Tổng quan

Trong chương này, khóa luận sẽ trình bày động lực nghiên cứu, phát biểu bài toán, thách thức của bài toán cũng như những đóng góp của nghiên cứu.

1.1 Động lực nghiên cứu

Nhận biết được hai ý nghĩa vô cùng to lớn của đề tài cả trong khoa học lẫn thực tiễn đã tạo động lực cho nhóm nghiên cứu thực hiện đề tài Phát hiện đối tượng trọng yếu.

1.1.1 Ý nghĩa khoa học

Phát hiện đối tượng trọng yếu có nguồn gốc từ khoa học nhận thức nhằm mô phỏng sát nhất cơ chế quan sát, tập trung nhận biết những vật thể nổi bật trước mắt của con người (Visual Attention). Việc nghiên cứu đề tài sẽ giúp ích rất lớn trong việc tái tạo hệ thống thị giác của con người, hỗ trợ rất lớn cho các tác vụ phức tạp hơn như tái nhận diện người, tự động chú thích ảnh, phân đoạn ngữ nghĩa, ... Từ đó có thể xây dựng các máy móc có khả năng quan sát, nhận biết và hành động ngày càng giống con người.

1.1.2 Ý nghĩa thực tiễn

Với sự bùng nổ của mạng học sâu cùng nguồn dữ liệu hình ảnh cực lớn đã tạo ra những mô hình thông minh thực hiện các tác vụ phức tạp như giám sát vật thể, phân đoạn hình ảnh, tái nhận diện người, nén ảnh, tự động chú thích ảnh, phát hiện sự kiện trong ảnh, hay các tác vụ trong đồ họa máy tính. Những tác vụ đó được ứng dụng rộng rãi trong nhiều lĩnh vực về an ninh, thương mại và điều khiển tự động nhằm tăng chất lượng cuộc sống, giảm sức lao động của con người đi đáng kể. Bởi vì phát hiện đối tượng trọng yếu được xem như là bước tiền xử lý cho các tác vụ

trên, nên việc có thể tạo nên một mô hình đạt được độ chính xác cao có ý nghĩa thực tiễn vô cùng to lớn. Đó cũng là kết quả mà khóa luận mong muốn đạt được.

1.2 Phát biểu bài toán

Phát hiện đối tượng trọng yếu nhận đầu vào là bức ảnh màu hai chiều, cho ra kết quả là một bức ảnh đen trắng chỉ giữ lại các vật thể quan trọng và bỏ đi các chi tiết không trọng yếu. Màu trắng thể hiện cho các vật thể chiếm sự chú ý, màu đen thể hiện cho các vật thể không trọng yếu hay phông nền bức ảnh. Nghiên cứu tập trung phát hiện vật thể trọng yếu không phân biệt lớp vật thể khác nhau cũng như số lượng. Xem **hình 1.2.1**.



Hình 1.2.1: Ảnh đầu vào (bên trái) và **nhãn** (bên phải) của bài toán phát hiện đối tượng trọng yếu trên ảnh màu hai chiều. Hai bức ảnh này được lấy từ tập **DUTS-TR** cho việc huấn luyện mô hình ở chương 3.

Khóa luận tập trung vào phát hiện một hoặc nhiều đối tượng trọng yếu, có thể thuộc nhiều lớp khác nhau trên ảnh màu hai chiều, không sử dụng các thông tin về độ sâu của ảnh.

1.3 Thách thức của bài toán

1.3.1 Định nghĩa đối tượng trọng yếu và cách thức đánh nhãn

Vật thể trọng yếu là như thế nào và cơ chế Visual Attention hoạt động ra sao là hai câu hỏi chưa có lời giải. Hơn nữa việc nhận biết nhiều đối tượng trọng yếu trong ảnh chứa nhiều vật thể thay vì chỉ nhận biết một đối tượng trong ảnh chứa duy nhất nó lại càng khó khăn hơn. Với những bức ảnh chỉ có một đối tượng (bộ dữ liệu **MSRA10K, ASD**) thì con người dễ dàng nhận biết và đồng tình chúng là trọng yếu. Nhưng một bức ảnh, khung hình bao gồm nhiều đối tượng thì việc nhìn nhận một đối tượng là quan trọng sẽ khác nhau giữa người với người. Ví dụ như những người có xu hướng nhìn tổng quan trước, chi tiết sau sẽ xem các vật thể lớn là trọng yếu, và ngược lại. Hay việc người đứng ở giữa thu hút sự chú ý nhất chiếm tỉ lệ cao nhưng không phải là hoàn toàn, sẽ có những người lại nhìn về một phía trước khi nhìn vào trung tâm.

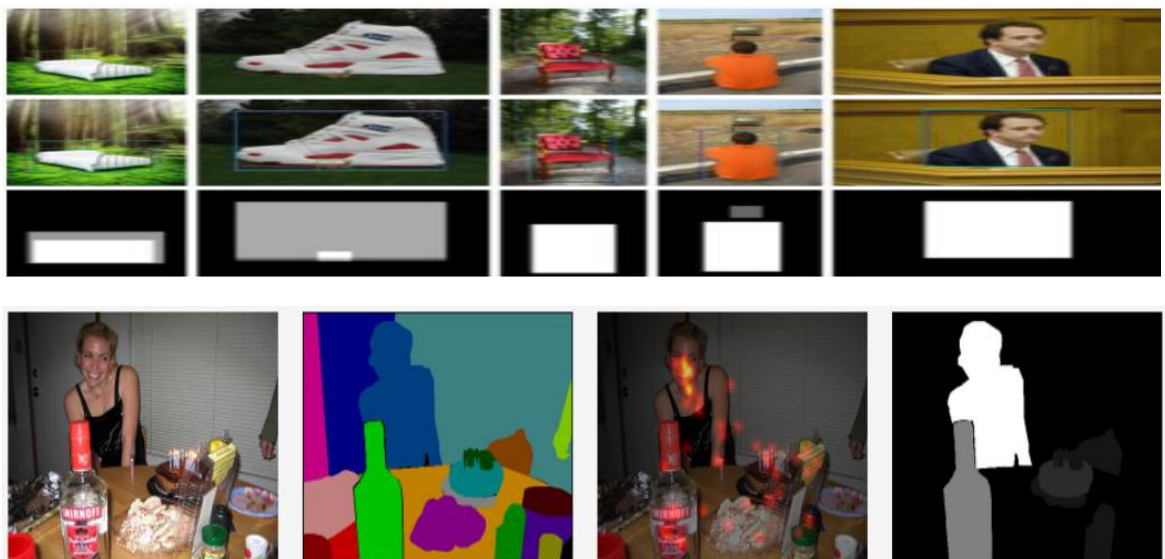
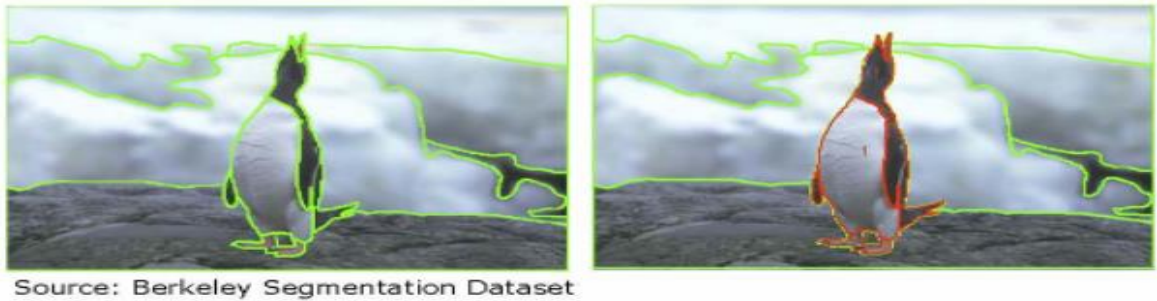
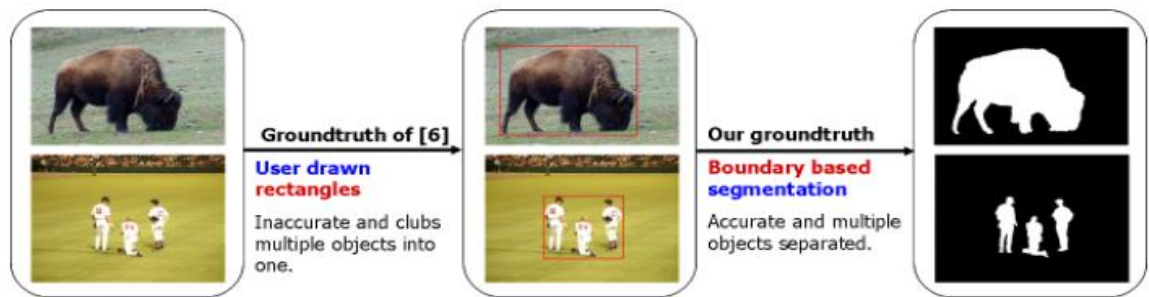
Do bài toán mang tính chất chủ quan như vậy, việc tạo ra dữ liệu chất lượng là phụ thuộc hoàn toàn vào khả năng nhận thức của người đánh nhãn. Sử dụng nhiều người đánh nhãn và chọn vật thể trọng yếu dựa trên tính nhất quán giữa những người đó sẽ giảm đi một phần nào đó tính chủ quan. Cách đánh nhãn của các tập dữ liệu chuẩn hiện này xem chi tiết tại **bảng 1.3.1** và **hình 1.3.1**. Tuy nhiên tính thiếu nhất quán trong các tập dữ liệu vẫn tồn tại **hình 1.3.2**.

Mặc dù vậy, các tập dữ liệu trên vẫn được sử dụng để huấn luyện (**DUTS-TR**) và đánh giá trong hầu hết các nghiên cứu cùng đề tài. Lý do đơn giản là tính thiếu nhất quán đó rõ ràng cũng tồn tại trong chính con người chúng ta. Hơn thế nữa cơ chế đánh nhãn sao cho thống nhất dựa trên các nghiên cứu khoa học nhận thức vẫn chưa được xác định.

Bộ dữ liệu	Thông tin	Cách thức đánh nhãn
ASD (2009)	Gồm 1000 bức ảnh và nhãn. Từng bức ảnh chỉ có một vật thể thu hút sự quan sát nhất.	Bức ảnh đã được người dùng vẽ hình chữ nhật bao quanh vật thể. Tác giả chỉ phân đoạn vật thể trọng yếu trong hình chữ nhật đó. Hình 1.3.1 dòng 1.
SOD (2010)	Gồm 300 bức ảnh và nhãn. Nhiều bức ảnh có trên một vật thể trọng yếu tương đồng với phần nền hoặc chạm biên cạnh ảnh.	Bộ dữ liệu đã được phân vùng/đoạn sẵn. 7 Người đánh nhãn cho vật thể theo họ là trọng yếu bằng cách nhấp chuột vào vùng ảnh đó. Các vật thể trọng yếu nhất quán giữa những người đánh nhãn trên ngưỡng được giữ lại. Ngược lại thì không. Hình 1.3.1 dòng 2.
MSRA10K (2015)	Giống ASD nhưng gồm 10,000 bức ảnh và nhãn. Được sử dụng để huấn luyện các mô hình học sâu trước 2017.	Giống ASD .
ECSSD (2015)	Gồm 1,000 bức ảnh và nhãn nhưng với cấu trúc nội dung bức ảnh phức tạp hơn.	5 người được yêu cầu đánh nhãn.
DUT-OMRON (2013)	Gồm 5,168 ảnh và nhãn với nền phức tạp và nội dung đa dạng.	25 người được yêu cầu đánh nhãn nhưng chỉ 5 người đánh nhãn cho cùng 1 bức ảnh. Vẽ bounding box trước, vật thể trọng yếu được giữ lại nếu có hơn k/5 người chọn. Sau đó mới đánh nhãn.

		Hình 1.3.1 dòng 3.
PASCAL-S (2014)	Gồm 850 ảnh và nhãn. Được xem là tập dữ liệu khách quan hơn các tập dữ liệu khác do cách đánh nhãn.	Gồm 8 người. Từng người sẽ được xem từng ảnh trong vòng 3 giây. Mắt của người đó được theo dõi bằng thiết bị. Các vật thể trọng yếu được quyết định bởi số điểm thiết bị nhận biết được trên vật thể của tất cả người tham gia. Hình 1.3.1 dòng 4.
HKU-IS (2015)	Gồm 4,447 ảnh phức tạp, chứa các vật thể trọng yếu không có sự liên kết, vật thể có bề ngoài giống với nền xung quanh. Ít nhất một vật thể trọng yếu chạm biên ảnh.	7,320 ảnh được thu thập và chọn lọc sao cho phức tạp như đã nói ở cột thông tin. Để giảm tính không nhất quán, 3 người được yêu cầu đánh nhãn cho tất cả bức ảnh trong khoảng hơn 3 tháng. Sử dụng tỉ lệ số điểm ảnh được đánh nhãn là trọng yếu của tất cả 3 người chia cho tổng số điểm ảnh được ít nhất 1 người xem là trọng yếu. Vứt bỏ những ảnh không có vật thể nào trên ngưỡng 0.9, chỉ còn lại 4,447 ảnh.
DUTS (2017)	Gồm 10,553 ảnh huấn luyện và 5,019 ảnh kiểm thử. Từ 2017 các mô hình học sâu sử dụng tập dữ liệu này để huấn luyện.	Các bức ảnh được thu thập từ tập dữ liệu ImageNetDET và SUN. Được đánh nhãn bởi 50 người.

Bảng 1.3.1: Cách thức đánh nhãn cho các bộ dữ liệu chuẩn được sử dụng phổ biến rộng rãi nhất hiện nay.



Hình 1.3.1: Cách thức đánh nhãn được thực hiện trong các tập dữ liệu chuẩn dòng 1 ASD, dòng 2 SOD, dòng 3 DUT-OMRON, dòng 4 PASCAL-S.



Hình 1.3.2: Những ví dụ về việc đánh nhãn thiếu sự nhất quán. Tất cả em bé trong ảnh **dòng 1** đều được đánh nhãn, nhưng những người trong ảnh **dòng 2** thì không. Ảnh **dòng 3**, cả con bướm và bông hoa đều được xem là trọng yếu nhưng ảnh **dòng 4** phải thì không.




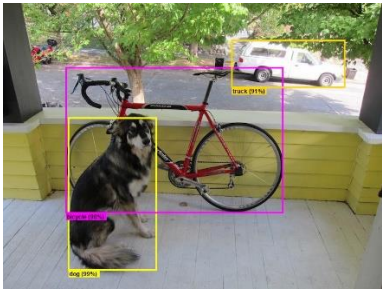
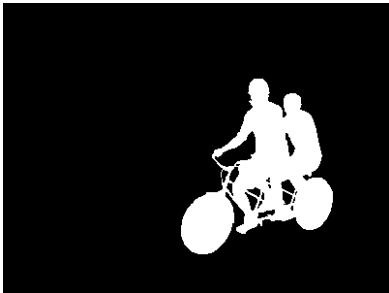

1.3.2 Tích hợp đặc trưng

Những bài toán dự đoán chính xác trên từng điểm ảnh như phân đoạn ngữ nghĩa, phát hiện đối tượng trọng yếu yêu cầu việc tích hợp các đặc trưng sao cho vẫn giữ được nguyên hai chiều về mặt không gian giống ảnh đầu vào phải thật sự hiệu quả. Tuy các bài báo từ năm 2017 đến nay đã đề xuất nhiều mô hình để giải quyết khó khăn trên nhưng lại đặt thêm một vấn đề mới là sự suy giảm của đặc trưng khi tích hợp (**leaping feature fusion operation**) [12] do sự cách biệt mức độ thông tin và độ phân giải.

1.3.3 Áp dụng các phương pháp từ các lĩnh vực tương tự

Phát hiện đối tượng trọng yếu là một nhánh nghiên cứu của phát hiện đối tượng—một tác vụ đã được nghiên cứu, phát triển và đạt được những kết quả vô cùng ấn tượng nhưng giữa chúng vẫn có sự khác nhau (**bảng 1.3.2**) dẫn đến việc áp dụng các kỹ thuật, phương pháp từ phát hiện đối tượng sang phát hiện đối tượng trọng yếu và ngược lại là một thách thức. Chẳng hạn như sau khi phát hiện được các vật thể nhưng việc xác định đối tượng nào trọng yếu giữa chúng là không dễ dàng (phải gán ngưỡng để quyết định vật thể nào là trọng yếu, không phù thuộc cho các tính huống phức tạp).

Thoạt nhìn bài toán phân đoạn ảnh và phát hiện đối tượng trọng yếu khá tương đồng. Tuy nhiên một cách biệt khiến các nghiên cứu về đề tài phân đoạn ảnh khó áp dụng thành công vào đề tài phát hiện đối tượng trọng yếu. Phân đoạn ảnh chính xác mà không cần nhiều thông tin toàn cục (cho kết quả tốt khi phân đoạn một phần bức ảnh). Trong khi để xác định vật thể quan trọng của bức ảnh, việc có cái nhìn tổng quan cả bức ảnh là vô cùng cần thiết. Do đó vùng nhìn thấy của nơ-ron (receptive field) của phát hiện đối tượng trọng yếu yêu cầu phải lớn hơn so với phân đoạn ảnh để có thể đạt kết quả chính xác.

	Phát hiện đối tượng	Phát hiện đối tượng trọng yếu	Phân đoạn/vùng ảnh
Đầu vào			
Đầu ra	<p>Tọa độ của bounding boxes và class scores cho từng bounding box.</p> 	<p>Một binary mask kích thước bằng ảnh đầu mà giá trị của từng pixel được gán nhãn thuộc các vật thể trọng yếu (không phân biệt sự khác nhau giữa các vật thể) hoặc background.</p> 	<p>Một ma trận mask mà giá trị của từng pixel được gán nhãn tương ứng với từng lớp vật thể.</p> 

Bảng 1.3.2: Sự khác nhau giữa các tác vụ phát hiện đối tượng, phát hiện đối tượng trọng yếu và phân đoạn/vùng ảnh.

1.4 Hướng tiếp cận

Với các thách thức trình bày ở trên, nhóm nghiên cứu sẽ lựa chọn mô hình học sâu Pyramidal Feature Shrinking Network (**PFSNet**) [12] kết hợp với Object-Contextual Representations (**OCR**) [11]. Khóa luận chọn PFSNet là vì mô hình chỉ tích hợp các khối đặc trưng liền kề, với các đặc trưng của chúng có sự cách biệt nhỏ cũng như mô hình đạt được kết quả tốt trong tác vụ này. OCR được lựa chọn là vì cơ chế self-attention hoạt động hiệu quả trên mô hình của bài toán phân đoạn ngữ nghĩa. Với OCR, mức độ trọng yếu của điểm ảnh sẽ được điều chỉnh dựa vào mức độ trọng yếu của tất cả các điểm ảnh khác.

Nhận thấy hàm mất mát Binary Cross Entropy (**BCE**) xem tầm quan trọng trong việc dự đoán tất cả điểm ảnh là như nhau, nhóm áp dụng hàm lỗi Adaptive Pixel Intensity (**API**) [16]. Hàm này đánh mạnh sai số của các điểm ảnh thuộc biên cạnh đối tượng trọng yếu, nơi ranh giới trọng yếu và không trọng yếu là sát sao.

1.5 Đóng góp

Dựa vào những kiến thức đã tìm hiểu từ các mô hình phát hiện vật thể đối tượng trọng yếu hiện nay cũng như các phương pháp sử dụng trong phân đoạn ảnh, khóa luận có những đóng góp sau đây:

- Phân tích sự hiệu quả của mô hình PFSNet.
- Xây dựng mô hình đề xuất dựa trên sự kết hợp mô hình PFSNet và Object-Contextual Representation.
- Áp dụng hàm lỗi Adaptive Pixel Intensity.
- So sánh kết quả của mô hình đề xuất với các mô hình hiện nay trên các tập dữ liệu chuẩn.

Chương 2

Các nghiên cứu liên quan

Chương này phân tích tiến trình phát triển của các mô hình học sâu và tập trung vào các phương pháp sử dụng để tích hợp đặc trưng.

2.1 Các nghiên cứu liên quan

Nhìn chung các mô hình học sâu phát hiện đối tượng quan trọng gồm những giai đoạn rút trích đặc trưng, tích hợp đặc trưng, phân lớp trên từng điểm ảnh, hậu xử lý. Tuy nhiên gần đây để giảm bớt tính toán cho mô hình, bước hậu xử lý đã được lược bỏ. Thông tin tổng quan xem **bảng thống kê 2.1.1**, chi tiết **bảng 2.1.2**.

Năm	Tên công trình.	Hội nghị	Kiến trúc	Backbone	Hàm lỗi	Tập dữ liệu huấn luyện
2015	Multi-Context Deep Learning (MCDL) [10]	IJCV	MLP+ super-pixel	GoogleNet	BCE	MSRA10K
	(Local Estimation and Global Search) LEGS [9]	CVPR	MLP+ segment		BCE+L2 regularization.	MSRA-B+PASCAL-S

2016	Encoded Low-level Distance Map and High-level Features (ELD) [8]	CVPR	MLP+ super-pixel	VGGNet	BCE	MSRA10K
	Deep Hierarchical Saliency Network (DHSNet) [6]	CVPR	FCN	VGGNet	BCE	MSRA10K+ DUT-OMRON
	Deep Contrast Learning (DCL) [5]	CVPR	FCN+ MLP+ super-pixel	VGGNet	Balanced BCE	MSRA-B
2017	Deeply Supervised Salient Object Detection (DSS) [7]	CVPR	FCN	VGGNet	Balanced BCE	MSRA-B+HKU-IS

	Aggregating Multi-level Convolutional Features for SOD (Amulet) [1]	ICCV	FCN	VGGNet	Balanced BCE	MSRA10K
2018	A Bi-directional Message Passing Model for SOD (BDMP) [3]	CVPR	FCN	VGGNet	Balanced BCE	DUTS
	Pixel-wise Contextual Attention Network (PiCANet) [13]	CVPR	FCN	VGGNet/ResNet-50	Balanced BCE	DUTS
2019	Boundary-Aware SOD Network (BASNet)	CVPR	FCN	ResNet-34	Bao gồm 3 hàm mất mát có trọng số ngang	DUTS

	[2]				nhau: BCE, hàm mất mát SSIM và IoU.	
	Cascaded Partial Decoder (CPD) [4]	CVPR	FCN	ResNet-50	BCE	DUTS
2021	Visual Saliency Transformer (VST) [15]	ICCV	Tokens-to-Tokens ViT	T2T ViT	BCE	DUTS
	Pyramidal Feature Shrinking for SOD (PFSNet) [12]	AAAI	FCN	ResNet-50	BCE và hàm lỗi IoU.	DUTS

2022	Extreme Attention Guided Salient Object Tracing Network (TRACER) [16]	AAAI	FCN	EfficientNet	Adaptive Pixel Intensity (API) bao gồm các hàm lỗi BCE, IoU, L1 nhưng có trọng số trên biên cạnh vật thể.	DUTS
------	---	------	-----	--------------	---	------

Bảng 2.1.1: Tổng quan các nghiên cứu liên quan.

Năm	Công trình.	Ý tưởng chính	Ưu điểm	Nhược điểm
2015	MCDL[10]	Hai cửa sổ có trung tâm là super-pixel, một cái lấy ra vùng ảnh cục bộ, một cái bao gồm cả bức ảnh. Hai ảnh này được cho qua GoogleNet để rút trích đặc trưng cục bộ và toàn cục tương ứng. Các đặc trưng này		Số lượng tham số lớn đến từ kiến trúc backbone phức tạp và MLP. Tính toán lặp lại nhiều lần do dự đoán cho từng super-pixel.

2016		được cho qua MLP để dự đoán mức độ quan trọng của super-pixel.		
	LEGS[9]	Một mạng nơ-ron MLP dự đoán bản đồ điểm quan trọng cục bộ. Đề xuất ứng viên kết hợp với saliency map cục bộ để tạo ra các véc-tơ đặc trưng toàn cục. Các đặc trưng này đi qua một MLP khác để dự đoán kết quả cuối cùng.		Sử dụng mạng nơ-ron nhiều lớp MLP. Saliency map cục bộ phải duyệt từng cửa sổ trên ảnh đầu vào. Dự đoán cho từng ứng viên.
	ELD[8]	Kết hợp các đặc trưng được rút trích bằng phương pháp truyền thông và bằng mạng tích chập.	Các đặc trưng truyền thông được rút trích cho từng super-pixel. Đặc trưng từ mạng tích chập được rút trích duy nhất một lần trên toàn bộ bức ảnh.	Mạng nơ-ron nhiều lớp. Dự đoán cho từng super-pixel.
	DHSNet[6]	Thay thế MLP bằng mạng tích chập nhưng chưa	Dần loại bỏ MLP ra khỏi kiến trúc.	Lớp fully connected và RCLs khiến

2017		hoàn toàn.		độ phức tạp của mô hình lớn.
	DCL [5]	Kiến trúc gồm hai luồng (2 streams). Một luồng là FCN dự đoán saliency map. Một luồng MLP dự đoán độ quan trọng cho từng super-pixel.	Luồng MLP sử dụng Spatial Pooling để lấy đặc trưng từ luồng FCN.	Kiến trúc hai luồng phức tạp. Dự đoán cho từng super-pixel.
	DSS [7]	Với mỗi đặc trưng có độ phân giải khác nhau sẽ cho ra bản đồ điểm quan trọng tương ứng. Đặc trưng được rút ra từ khối tích chập trước sẽ được kết hợp với các đặc trưng khối tích chập sau thông qua nội suy song tính và các lớp chập (short connections). Giám sát sâu được áp dụng để hướng	Kiến trúc FCN	Tích hợp các đặc trưng có độ cách biệt cao về độ phân giải và mức độ thông tin qua nội suy tuyến tính x2, x4, x8 để tăng kích thước.

2018		dẫn mô hình huấn luyện.		
	Amulet [1]	Tất cả các khối đặc trưng (5 khối) được tích hợp trực tiếp tạo ra 5 khối đặc trưng với 5 độ phân giải gồm cả thông tin bậc cao và bậc thấp. Giám sát sâu được áp dụng	FCN	Phương pháp tích hợp đặc trưng vẫn đơn giản dẫn đến việc mất mát chi tiết của vật thể được phát hiện.
	BDMP [3]	Chọn lọc đặc trưng bằng cách truyền thông tin giữa các đặc trưng ban đầu theo hai chiều. Các thông tin về các vật thể không trọng yếu được giữ lại, chi tiết của vật thể quan trọng được truyền đi thông qua một cổng kiểm soát.	Cho kết quả tốt so với các mô hình cùng thời chứng tỏ việc tích hợp đặc trưng đã được cải thiện.	
	PiCANet [13]	PiCANet bao gồm hai mô-đun Global	Cơ chế mới (LSTM) chọn lọc đặc trưng	Mô-đun phức tạp.

		<p>PiCANet và Local PiCANet. Global PiCANet sử dụng LSTM hai chiều, chạy theo hai hướng dọc và ngang để chọn lọc đặc trưng toàn cục cho từng điểm ảnh. Local PiCANet tương tự nhưng thay vì LSTM thì sử dụng mạng tích chập để lọc ra thông tin cục bộ.</p>	<p>phức tạp cho kết quả tương đối tốt.</p>	
2019	BASNet [2]	<p>Một hàm mất mát hỗn hợp được tạo ra từ ba hàm mất mát cũ.</p>	<p>Hàm mất mát hỗn hợp có thể phát hiện vật thể trọng yếu chính xác tại biên cạnh lẫn bên trong chúng. Kết quả mô hình vượt trội lúc bấy giờ khi với các vật thể trọng yếu có biên cạnh vô cùng phức tạp được phát hiện vô cùng chính</p>	

			<p>xác.</p> <p>Là tiền đề cho các nghiên cứu sau cho ra các hàm lỗi đánh mạnh vào các điểm ảnh khó dự đoán như ở gần biên cạnh. Từ</p>	
	CPD [4]	<p>Nhận biết được các đặc trưng từ các khối tích chập đầu tiên (1, 2) gây nên độ phức tạp tính toán vô cùng lớn trong khi chỉ giúp tăng độ chính xác không đáng kể, tác giả hi sử dụng các đặc trưng từ khối tích chập 4, 5. Đánh đổi một phần độ chính xác để đạt tốc độ thực khi mà các đặc trưng được sử dụng có độ phân giải thấp.</p>	<p>Đạt được độ chính xác tương đồng với các mô hình lúc bấy giờ nhưng đạt tốc độ thực. FPS > 62.</p>	<p>Đánh đổi độ chính xác.</p>
2021	VST [15]	Kiến trúc Tokens-to-Tokens ViT để	<p>Phương pháp học đa tác vụ: phát hiện vật</p>	<p>Để có thể cải thiện thêm</p>

		<p>rút trích đặc trưng. Phương pháp upsample mới được sử dụng để đáp ứng bài toán dự đoán trên từng điểm ảnh: mô-đun reverse T2T.</p>	<p>thể trọng yếu và biên cạnh vật thể thông qua hai tokens t_s và t_b được thêm vào trong decoder. Sự hiệu quả của hai mô-đun T2T và reverse T2T đã cho kết quả thu được tốt nhất lúc bấy giờ. Hơn nữa kiến trúc này có thể đồng thời giải quyết bài toán RGB-D SOD khi được cung cấp thêm dữ liệu theo chiều sâu.</p>	<p>mô hình, đòi hỏi một kiến trúc decoder mới.</p>
	PFSNet [12]	<p>Chỉ tích hợp các đặc trưng liên kế nhau, tạo ra một đặc trưng có độ phân giải bằng với ảnh đầu vào nhưng không sử dụng tăng kích thước đặc trưng quá nhiều lần.</p>	<p>Đạt được đặc trưng có độ phân giải cao. Tránh được hiện tượng suy giảm đặc trưng khi tích hợp.</p>	<p>Mô hình vẫn nhằm lẫn giữa các vật thể không trọng yếu.</p>

2022	TRACER ^[16]	Sử dụng backbone EfficientNet ít tham số hơn so với các kiến trúc rút trích khác. Sử dụng fast Fourier transform và hàm lỗi biên cạnh (cũng xài hàm Adaptive Pixel Intensity, nhưng áp dụng cho biên cạnh) để giữ lại cũng như tăng cường thông tin biên cạnh. Ba mô-đun attention được sử dụng để chọn lọc và tăng cường đặc trưng: Masked edge attention, Union attention, Object attention).	Với những cải tiến mới và mô hình nhẹ EfficientNet, mô hình có kết quả tốt nhất hiện tại theo bảng xếp hạng trên trang paperswithcode.com. Hơn thế nữa, từ EfficientNet7 xuống EfficientNet0 mô hình có độ phức tạp giảm theo độ chính xác nhưng tốc độ tăng do lượng tham số giảm mạnh.	
------	------------------------	---	--	--

Bảng 2.1.2: Phân tích ưu nhược điểm của các công trình liên quan.

2.2 Phân tích hướng phát triển và cải tiến

Từ những phân tích các nghiên cứu liên quan ở bảng trên, nhóm nhận thấy được các xu hướng cải tiến trong mạng tích chập:

- Thay thế hoàn toàn các mô hình có sử dụng phương pháp đề xuất ứng viên, super-pixels, hậu xử lý bằng các mô hình học sâu một giai đoạn duy nhất có thể huấn luyện từ đầu đến cuối. Việc này nhằm tăng tốc độ cũng như độ chính xác của bài toán khi mà các phương pháp đi trước không thể hiện tốt trong những trường hợp phức tạp.
- Sự lựa chọn các backbone của mô hình để rút trích đặc trưng của ảnh chuyển từ các kiến trúc nhiều tham số như VGG16 sang các kiến trúc nhẹ hơn, tốt hơn như Resnet-34, Resnet-50, EfficientNet.
- Tích hợp đặc trưng sao cho hiệu quả nhất, có chọn lọc, kiểm soát bằng các cơ chế spatial attention và channel attention. Tránh sử dụng các phương pháp phức tạp như tích chập LSTM, GRU.
- Phát hiện đối tượng trọng yếu với các kích thước khác nhau ngày càng được quan tâm.
- Hàm lỗi không còn đơn giản như Binary Cross Entropy (BCE) nữa. Nhận thấy được độ khó trong việc phân lớp là khác nhau giữa các điểm ảnh (các điểm ảnh ở biên cạnh vật thể là khó hơn so với các điểm ảnh khác), các hàm lỗi phức tạp như API đặt nhiều trọng số hơn vào các điểm ảnh đó thông qua các cơ chế như Erosion-Dilation, bộ lọc đa kích thước.

Vài năm trở lại đây, kiến trúc Vision Transformer (ViT) đã cho thấy sự vượt trội của chúng so với các mô hình sử dụng mạng tích chập khi được áp dụng vào các tác vụ của thị giác máy tính. Lĩnh vực phát hiện đối tượng trọng yếu (cụ thể là trên ảnh màu hai chiều) cũng đã đón nhận một mô hình **Visual Saliency Transformer**[\[15\]](#) vào năm 2021 và đạt được kết quả khá tốt. Tuy nhiên các nghiên cứu sử dụng kiến

trúc Transformer cho đề tài này vẫn còn rất ít. Hơn thế nữa việc có thể tạo ra một decoder có khả năng khôi phục đặc trưng về không gian ảnh là một thử thách lớn.

2.3 Kết luận

Trong chương này, nhóm đã đề cập và phân tích các nghiên cứu theo nhóm là đóng vai trò quan trọng trong việc đưa ra các hướng phát triển đến hiện nay.

Với những phân tích và lý do đề cập ở trên, khóa luận sẽ tiếp tục nghiên cứu các mô hình sử dụng tích chập, cụ thể là mô hình PFSNet để rút trích cũng như tích hợp các đặc trưng có độ phân giải cao. Đồng thời sẽ áp dụng mô-đun Object-Contextual Representations từ mô hình phân đoạn hình ảnh tốt nhất hiện nay để học ra hai véc-tơ đại diện cho độ trọng yếu. Thử nghiệm này hi vọng sẽ có thể tăng độ chính xác cho mô hình PFSNet.

Trong chương tiếp theo, khóa luận sẽ đưa ra hướng tiếp cận với chi tiết các giai đoạn rút trích đặc trưng, tích hợp đặc trưng, phân lớp và hàm lỗi. Ở mỗi giai đoạn, khóa luận sẽ đưa ra các lý do cho từng lựa chọn mà theo nhóm hi vọng sẽ tăng độ chính xác của mô hình.

Chương 3

Phương pháp đề xuất

Mô hình được đề xuất nhìn chung có ba quá trình chính: rút trích đặc trưng, chọn lọc và tích hợp đặc trưng, tính toán hàm lỗi. Mô hình là một mạng tích chập thống nhất, có thể huấn luyện từ đầu đến cuối và không sử dụng bất cứ kỹ thuật hậu xử lý nào.

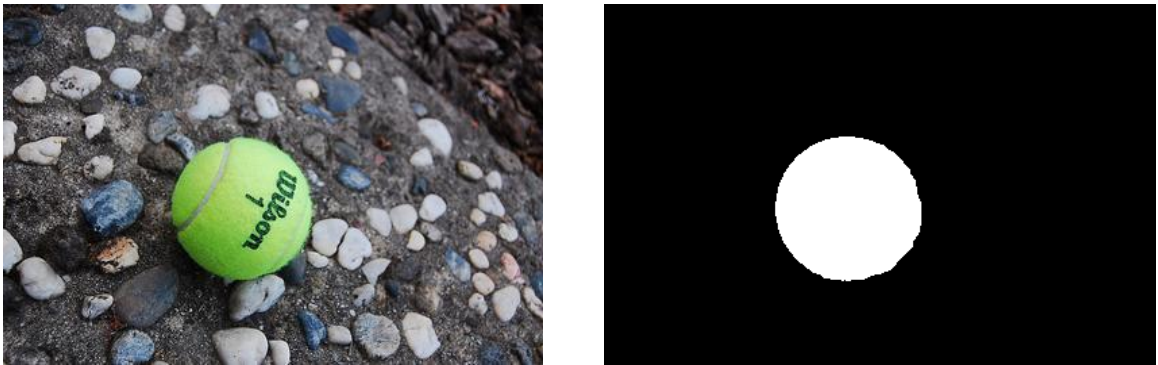
Trong chương này, nhóm sẽ trình bày chi tiết từng quá trình trên cũng như đưa ra lý do cho sự lựa chọn đó có thể mang đến kết quả hứa hẹn cho mô hình.

3.1 Rút trích đặc trưng

Các đối tượng được xem là trọng yếu trong ảnh khi chúng vừa có sự nổi bật cục bộ so với các điểm ảnh lân cận (ví dụ như màu sắc), vừa chiếm sự chú ý của toàn bộ bức ảnh so với các đối tượng khác (ví dụ như vị trí, kích thước của vật thể, khoảng cách của chúng so với các vật thể khác). Vì thế, việc rút trích được các đặc trưng đa bậc cho bài toán phát hiện đối tượng trọng yếu là vô cùng cần thiết.

Các kiến trúc mạng nơ-ron tích chập như Resnet, VGG cho thấy khả năng rút trích đặc trưng của ảnh vô cùng hiệu quả trong đa số các tác vụ của Thị giác máy tính. Khi bức ảnh đi qua các lớp tích chập đầu tiên, mô hình rút ra được các đặc trưng bậc thấp, giàu thông tin cục bộ (như biên cạnh, đường nét của vật thể). Đến các khối tích chập ở giữa, mô hình học được hình dạng trung gian như mắt, tai, bánh xe. Khi qua các khối tích chập cuối cùng, đặc trưng phức tạp của toàn bức ảnh góp phần rất lớn trong việc dự đoán chính xác, ví dụ trong tác vụ phân lớp hình ảnh. Càng đi sâu qua các khối tích chập và max pooling, receptive field (vùng ảnh mà một nơ-ron có thể nhìn thấy) tăng dần, mô hình có thể học được các đặc trưng phức tạp, súc tích, giàu ngữ nghĩa và toàn cục của bức ảnh. Kết hợp các đặc trưng bậc cao, độ phân giải thấp với các đặc trưng bậc thấp, độ phân giải cao trước đó sẽ có

thể cung cấp thông tin hữu ích cho từng điểm ảnh để xác định điểm ảnh đó có trọng yếu hay không.



Hình 3.1.1: Các viên sỏi trong ảnh bên trái có độ tương phản so với các điểm ảnh xung nhưng chúng không được xem là trọng yếu. Để phát hiện vật thể trọng yếu thì yêu cầu phải rút ra được các đặc trưng vừa mang thông tin cục bộ, vừa mang thông tin toàn cục, giàu ngữ nghĩa của bức ảnh.

Trong luận văn này, mô hình Resnet-50 (**hình 3.1.1**) được chọn để làm backbone cho mô hình vì có thể rút trích ra các đặc trưng đa bậc. Từ đó khi tích hợp các đặc trưng đa bậc có thể rút trích ra đặc trưng cho từng điểm ảnh. Mặc dù đặc trưng được rút trích có độ phân giải lớn nhất chỉ bằng nửa ảnh đầu vào, nhưng mô hình sâu, nhẹ, hiệu quả tốt đã khiến chúng dần thay thế VGG.

Bởi kiến trúc Resnet giảm độ phân giải của ảnh đầu vào đi nhanh chóng nhằm giảm số lượng các phép toán. Trong khi VGG thực hiện nhiều lớp tích chập trên các khối đặc trưng (feature map) có độ phân giải bằng ảnh đầu vào. Hơn nữa Resnet sử dụng các kết nối tắt để có thể huấn luyện mô hình sâu hơn so với VGG, dẫn đến các đặc trưng được rút trích có độ phức tạp cao hơn, giàu ngữ nghĩa hơn và vùng nhìn thấy của mỗi nơ-ron sẽ lớn hơn. Xem **bảng 3.1.1** tóm tắt so sánh hai kiến trúc. Từ những phân tích trên, mô hình sử dụng Resnet-50 có tốc độ huấn luyện, kiểm thử nhanh hơn và chính xác hơn.

	Resnet-50	VGG-16
Số lượng tham số	Hơn 23 triệu tham số. Nhẹ hơn.	138 triệu tham số
Số lớp	50 lớp – sâu hơn giúp học được các đặc trưng phức tạp hơn.	16 lớp
Kích thước đầu ra của ảnh khi đi qua các lớp tích chập đầu tiên	Giảm đi 4 lần so với kích thước ảnh đầu vào. Từ đó số lượng phép toán FLOPs.	Giảm đi 2 lần so với kích thước ảnh đầu vào

Bảng 3.1.1: So sánh kiến trúc Resnet-50 với VGG-16.

Gọi I là ảnh đầu vào có kích thước $H \times W \times 3$, khi qua các khối tích chập thu được các khối đặc trưng $\{L_i / i = 1, 2, \dots, 5\}$ có kích thước tương ứng $[\frac{H}{2^i}; \frac{W}{2^i}; C_i]$. Chi tiết kiến trúc Resnet-50, đặc trưng được rút ra từ lớp nào, tên và kích thước đầu ra của các đặc trưng được thể hiện trong **bảng 3.1.2**.

Tên khối tích chập	Resnet-50	Kích thước đầu ra	Tên đầu ra
Conv1	7x7, 64 stride 2	$\frac{H}{2} \times \frac{W}{2} \times 64$	L1
Conv2_x	3x3 maxpool, stride 2		L2
	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\frac{H}{4} \times \frac{W}{4} \times 256$	
Conv3_x	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$	$\frac{H}{8} \times \frac{W}{8} \times 512$	L3
Conv4_x	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$	$\frac{H}{16} \times \frac{W}{16} \times 1024$	L4
Conv5_x	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$	$\frac{H}{32} \times \frac{W}{32} \times 2048$	L5

Bảng 3.1.2: Một phần của kiến trúc **Resnet-50** được sử dụng để rút trích đặc trưng của ảnh đầu vào với kích thước $H \times W \times 3$ (lớp average pool và fully connected được vớt bỏ). Kết quả của quá trình thu được 5 khối đặc trưng như **L1**, **L2**, **L3**, **L4**, **L5** và kích thước đầu ra tương ứng

3.2 Chọn lọc và tích hợp đặc trưng

Các bài báo nghiên cứu về phát hiện đối tượng trọng yếu trong ảnh màu sử dụng mạng nơ-ron tích chập chủ yếu cố gắng tìm ra phương pháp kết hợp các khối đặc trưng đa bậc sao cho hiệu quả nhất. Đã được đề cập ở phần trước, các khối đặc trưng được rút trích ra từ backbones như VGG, Resnet có sự khác biệt vừa về mặt không gian (chiều dài, chiều rộng), vừa về mặt thông tin bức ảnh được rút trích. Sự khác biệt lớn nhất khi hai khối đặc trưng được rút trích ở đầu và cuối kiến trúc backbone.

Vì thế trong luận văn này, các mô-đun chỉ tích hợp các khối đặc trưng liên kế để giảm đi hiện tượng suy giảm đặc trưng do sự cách biệt giữa chúng. Hai mô-đun để tích hợp đặc trưng liên kế được sử dụng trong mô hình là Scale-aware Enrichment Module (SEM) và Adjacent Fusion Module (AFM) từ kiến trúc PFSNet.

3.2.1 Scale-aware Enrichment Module (SEM)

Các khối đặc trưng được rút trích từ các khối tích chập đầu và giữa của kiến trúc Resnet-50 chỉ học được các biên cạnh, hình dạng chung chung của bức ảnh, bất kể là vật thể đó có trọng yếu hay không. Do đó mô-đun SEM được đề xuất để truyền luồng thông tin phức tạp, toàn cục, giàu ngữ nghĩa **L5** đến các khối đặc trưng trước để chọn lọc lần một. Các biên cạnh, hình dạng của các vật thể không trọng yếu sẽ được bỏ đi, của vật thể trọng yếu được giữ lại.

Nhận biết được vùng nhìn thấy của nơ-ron theo lý thuyết nhỏ hơn nhiều so với thực tế, mô-đun sử dụng các lớp tích chập giãn nở (dilated convolutions) nhằm tăng vùng nhìn thấy. Từ đó mô hình có thể phát hiện các vật thể trọng yếu có kích thước khác nhau cũng như xác định điểm ảnh trọng yếu hiệu quả hơn.

Mô-đun cũng cho ra các khối đặc trưng có số lượng kênh (channels) nhỏ hơn nhiều so với các khối đặc trưng đầu vào (**L_i**). Từ đó giảm đi đáng kể số lượng tham số của mô hình. Xem **bảng 3.2.1**.

Cách thức hoạt động của mô-đun có thể được diễn giải như sau:

$$\mathbf{S}_5 = \mathbf{wcat}(\text{conv}_{\text{dilation}=1}(\mathbf{L}_5), \text{conv}_{\text{dilation}=2}(\mathbf{L}_5))$$

$$\mathbf{S}_i = \mathbf{wcat}(\text{conv}_{\text{dilation}=1}(\mathbf{L}_i), \text{conv}_{\text{dilation}=2}(\text{Up}(\text{conv}_{\text{dilation}=2}(\mathbf{L}_{i+1})) + \mathbf{L}_i)), i=1, \dots, 4$$

(3.2.1.1)

Với:

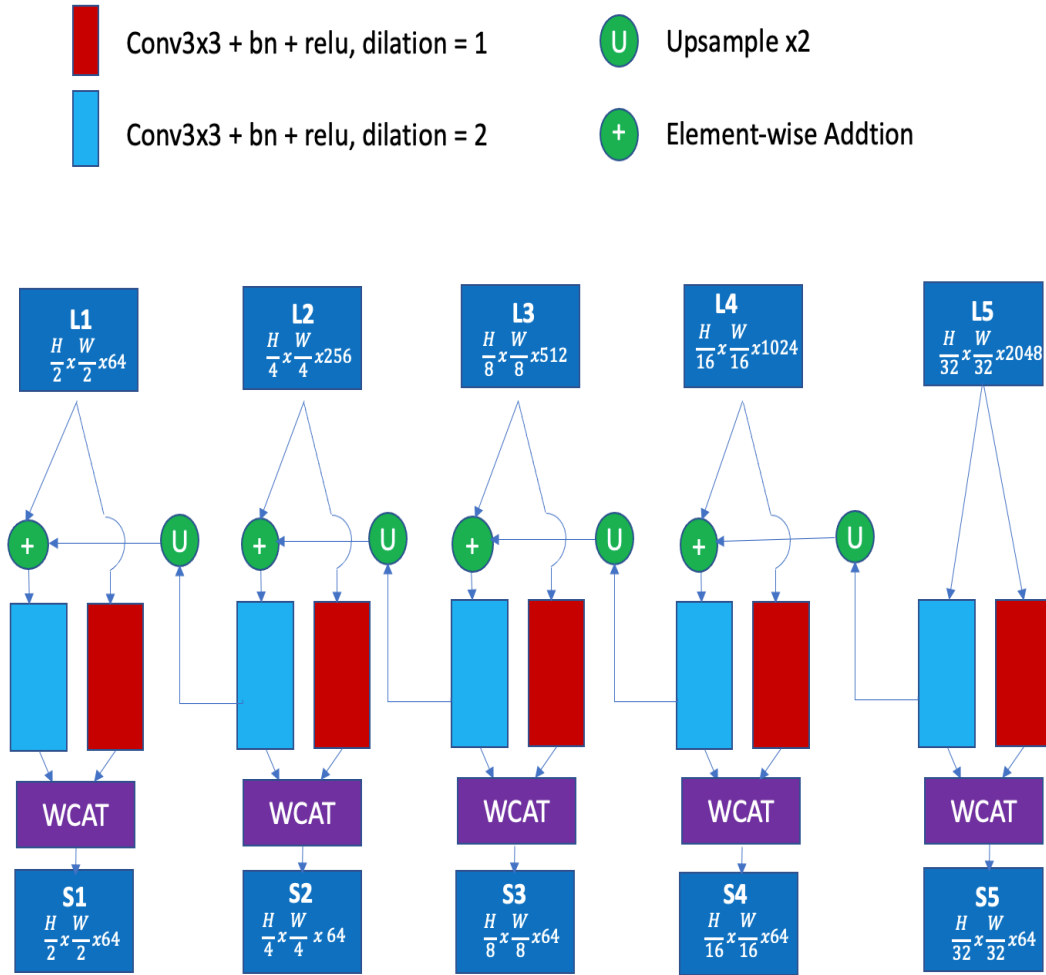
- $\text{conv}_{\text{dilation}=d}$ là một chuỗi các phép toán theo thứ tự tích chập giãn nở 3x3, Batch Normalization, ReLU.
- Up: Nội suy song tuyến.
- $+$: phép toán cộng từng phần tử của hai tensor.

- **wcat**: mô-đun này đóng vai trò là channel attention. Đầu vào của mô-đun này là hai khối đặc trưng đã được rút trích từ **Li** theo sau là các lớp tích chập giãn nở với hệ số giãn nở khác nhau. Mô-đun này sẽ chọn lọc channel (kênh) nào của khối đặc trưng là hữu ích cho việc dự đoán. Chi tiết ở mục **3.2.3**.

Xem **hình 3.2.1** minh họa cách thức hoạt động của mô-đun SEM. Kết quả của quá trình tích hợp đặc trưng ở giai đoạn này được thể hiện ở bảng dưới.

Tên khối đặc trưng	S1	S2	S3	S4	S5
Kích thước	$\frac{H}{2} \times \frac{W}{2} \times 64$	$\frac{H}{4} \times \frac{W}{4} \times 64$	$\frac{H}{8} \times \frac{W}{8} \times 64$	$\frac{H}{16} \times \frac{W}{16} \times 64$	$\frac{H}{32} \times \frac{W}{32} \times 64$

Bảng 3.2.1: Các khối đặc trưng và kích thước của chúng sau khi được tích hợp trong mô-đun SEM.



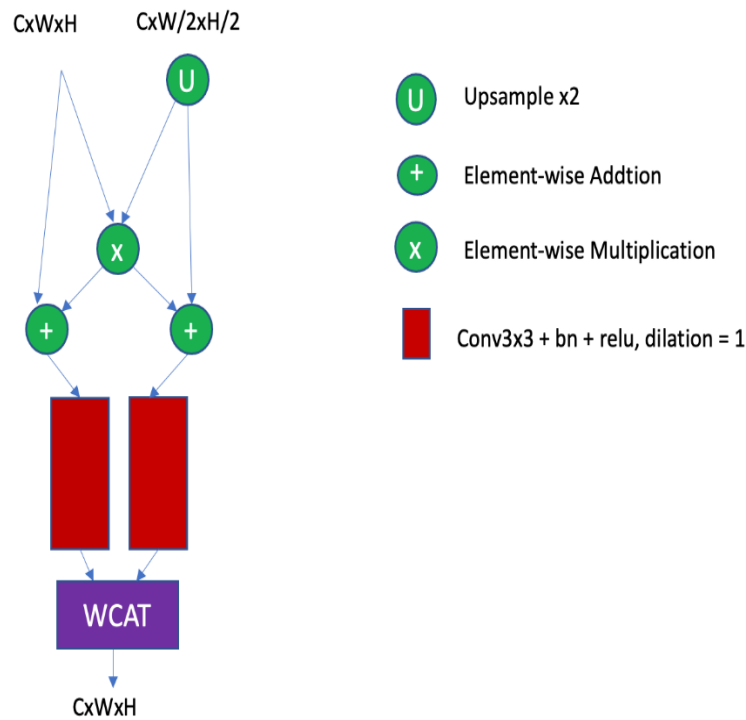
Hình 3.2.1: Quá trình tích hợp đặc trưng trong mô-đun SEM.

3.2.2 Adjacent Fusion Module (AFM)

Sau khi truyền luồng thông tin từ **L5** tới các khối đặc trưng **L_i** với $i=1, \dots, 4$ trong mô-đun SEM ở mục trên, việc xuất hiện nhiễu trong quá trình là không thể tránh khỏi. Mô-đun AFM đóng vai trò tích hợp các khối đặc trưng liền kề đôi một với nhau, giữ lại được những đặc trưng có sự tương đồng và lược bỏ nhiễu.

Nếu ta xem đầu vào hai khối đặc trưng liền kề của mô-đun AFM là khối đặc trưng cha và mẹ, thì đầu ra là khối đặc trưng được thừa hưởng các đặc tính có sự giống nhau giữa cha và mẹ, cũng như tốt nhất, phù hợp cho quá trình tích hợp tiếp theo hiệu quả hơn.

Hai khối đặc trưng cha, mẹ được lọc bỏ nhiễu, giữ lại đặc trưng tương đồng giữa chúng bằng phép nhân từng phần tử của hai tensor. Khối đặc trưng tương đồng này sẽ được cộng vào hai khối đặc trưng cha, mẹ để tăng cường thông tin hữu ích. Cuối cùng, khối đặc trưng con sẽ chọn lọc các đặc tính (các kênh) từ hai khối đặc trưng trên thông qua mô-đun **wcat** sẽ được đề cập ở mục sau. Chi tiết quá trình tích hợp của mô-đun AFM xem **hình 3.2.2** minh họa.

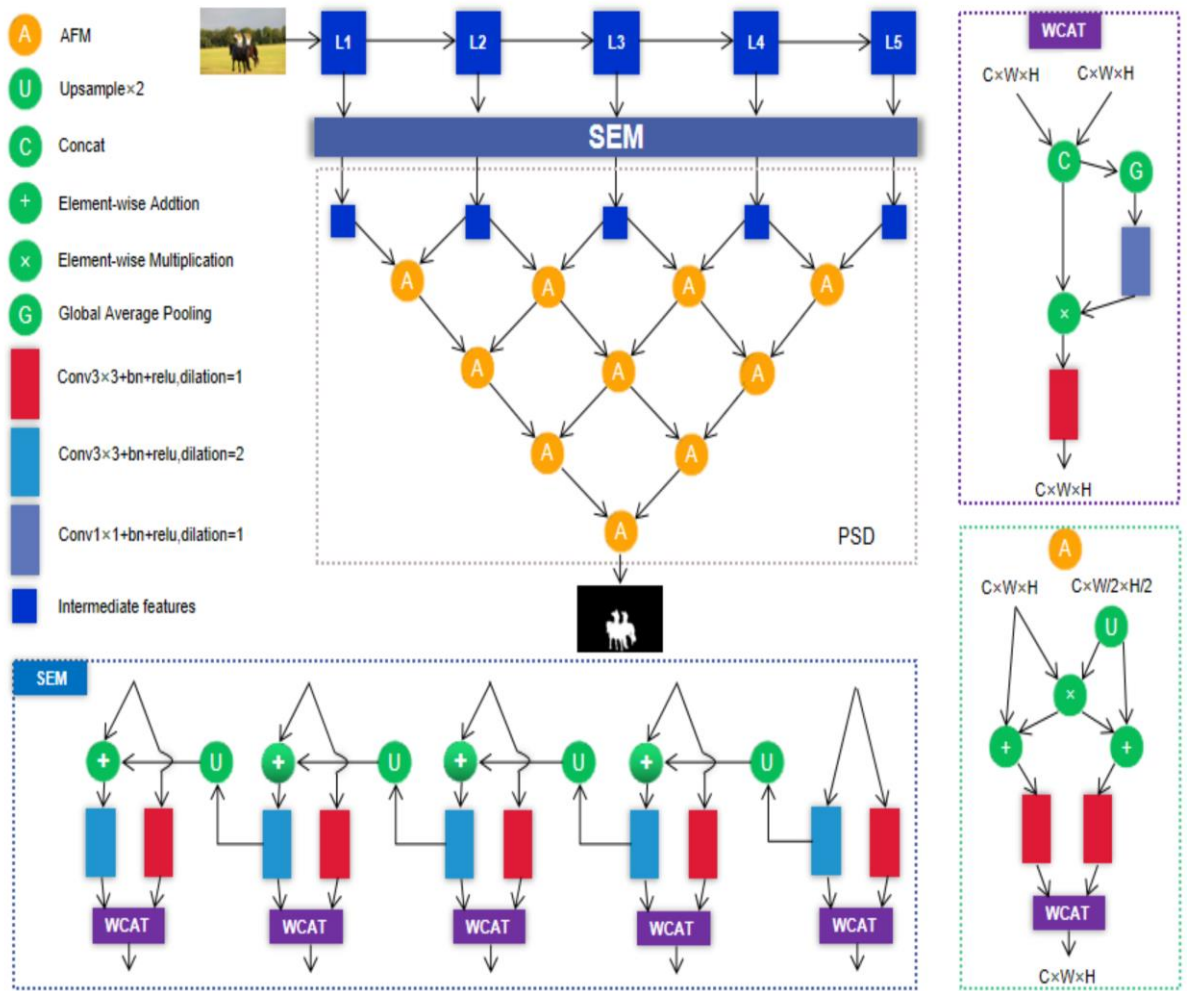


Hình 3.2.2: Quá trình tích hợp đặc trưng trong mô-đun AFM.

Quá trình tích hợp của các mô-đun AFM theo dạng hình chóp (**hình 3.2.3**), chiều cao, chiều rộng của khối đặc trưng con thu được bằng với khối đặc trưng đầu vào bên trái. Vì thế, kết quả của quá trình ta thu được khối đặc trưng có kích thước $\frac{H}{2} \times \frac{W}{2} \times 64$.

3.2.3 Cơ chế channel-attention (WCAT)

Trong mạng nơ-ron thông thường (**ANN**), các đặc trưng được rút trích ở lớp này sẽ là đầu vào của lớp tiếp theo. Đặc trưng được rút trích là một scalar (feature), tương ứng với một node trong mạng. Trong khi mạng tích chập (**CNN**) mỗi bộ lọc(filter) sẽ trượt trên khối đặc trưng, rút trích ra một hình dạng chung (pattern) trên toàn bộ điểm không gian (feature map). Từ đó mỗi bộ lọc thu được một kênh (channel). Vậy mỗi kênh của khối đặc trưng được rút trích trong **CNN** tương ứng với mỗi node trong **ANN**. Ví dụ bộ lọc đầu tiên của lớp tích chập ở giữa kiến trúc rút trích ra hình dạng tròn như bánh xe trên bức ảnh, bộ lọc thứ hai rút ra hình dạng vuông như màn hình. Cơ chế channel-attention sẽ chọn lọc ra các kênh (đặc trưng) phù hợp với từng bức ảnh đầu vào khác nhau bằng cách nhân trọng số kênh với các kênh tương ứng. Cơ chế channel-attention này được sử dụng khá phổ biến trong mạng tích chập. Mô-đun **WCAT** được cài đặt y chang và được sử dụng trong các mô-đun SEM và AFM (**hình 3.2.3**).



Hình 3.2.3: Kiến trúc PFSNet[12] để rút trích và tích hợp các đặc trưng đa bậc. Mô-đun WCAT đóng vai trò như một cơ chế channel attention, giúp chọn lọc đặc trưng. Phép toán với từng phần tử trong mô-đun nhằm nâng cao và tăng cường các đặc trưng tương đồng.

3.2.4 Cơ chế dựa trên self-attention (OCR)

Hình 3.2.3 cho thấy cái nhìn tổng quát cả quá trình tích hợp đặc trưng trên. Sau quá trình đó thu được khối đặc trưng hay các véc-tơ đặc trưng đại diện mức độ trọng yếu của từng điểm ảnh (với kích thước bằng một nửa ảnh đầu vào):

$$\mathbf{F}_{\text{interm}} = \{f_p^{\text{interm}} \in \mathbf{R}^{64} \mid p = 1, 2, \dots, \frac{H}{2} \times \frac{W}{2}\} \quad (3.2.4.1)$$

Quá trình phân lớp mức độ trọng yếu được thực hiện trên khối đặc trưng để thu được một bản đồ điểm quan trọng trung gian (intermediate saliency map):

$$\mathbf{Sal}_{\text{interm}} = \{Sal_p^{\text{interm}} \in [0, 1] \mid p = 1, 2, \dots, \frac{H}{2} \times \frac{W}{2}\}. \quad (3.2.4.2)$$

Bản đồ này được xem là trọng số, kết hợp với các đặc trưng điểm ảnh trong $\mathbf{F}_{\text{interm}}$ thu được một véc-tơ đặc trưng đại diện cho mức độ trọng yếu \mathbf{V}_{sal} :

$$\mathbf{V}_{\text{sal}} = \frac{\sum_{p=1}^{\frac{H}{2} \times \frac{W}{2}} Sal_p^{\text{interm}} f_p^{\text{interm}}}{\sum_{p=1}^{\frac{H}{2} \times \frac{W}{2}} Sal_p^{\text{interm}}} \quad (3.2.4.3)$$

Cơ chế self-attention nhận đầu vào là $\mathbf{F}_{\text{interm}}$ và \mathbf{V}_{sal} , với các véc-tơ điểm ảnh trong $\mathbf{F}_{\text{interm}}$ đóng vai trò là **query**, \mathbf{V}_{sal} đóng vai trò là **key** và **value**. Kết quả thu được là các véc-tơ đặc trưng điểm ảnh nằm trong khối đặc trưng \mathbf{X} được điều chỉnh dựa vào sự tương đồng của chúng với véc-tơ đại diện mức độ trọng yếu \mathbf{V}_{sal} . Khối đặc trưng \mathbf{X} này sẽ được nối với $\mathbf{F}_{\text{interm}}$ theo chiều kênh để cho ra khối đặc trưng cuối cùng $\mathbf{F}_{\text{final}}$. Khối đặc trưng này được dùng để dự đoán bản đồ điểm quan trọng cuối cùng $\mathbf{Sal}_{\text{final}}$. Mô-đun OCR cho thấy sự hiệu quả trong phân đoạn ngữ nghĩa, khi mà số key và value là n lớp vật thể cần được dự đoán. Nhận thấy khả năng có thể áp dụng vào bài toán phát hiện đối tượng trọng yếu, nhóm áp dụng mô-đun trên để nhằm cải tiến mô hình với số key và value là hai, một cho véc-tơ đại diện cho mức độ trọng yếu, hai là đại diện cho không trọng yếu. Tuy nhiên nhận thấy véc-tơ

thứ hai thực chất chỉ là phần bù, nhóm chỉ rút ra véc-tơ đặc trưng đại diện cho mức độ trọng yếu. Xem **bảng 3.2.2** thể hiện khả năng áp dụng OCR từ phân đoạn ngữ nghĩa cho phát hiện đối tượng trọng yếu.

	Phân đoạn ngữ nghĩa	Phát hiện đối tượng trọng yếu
Phân lớp	Phân lớp n vật thể	Phân lớp các đối tượng trọng yếu, không trọng yếu.
Số lượng véc-tơ đặc trưng đại diện	N véc-tơ đại diện cho n lớp vật thể.	1 véc-tơ đại diện cho mức độ trọng yếu (véc-tơ đại diện cho không trọng yếu chỉ là phần bù)

Bảng 3.2.2: Áp dụng OCR từ phân đoạn ngữ nghĩa cho bài toán phát hiện đối tượng trọng yếu.

Chi tiết cơ chế OCR được trình bày như sau:

$$\begin{aligned}
 \mathbf{key} &= \text{ReLU}(\text{BatchNorm}(\text{Conv}_k(\mathbf{V}_{\text{sal}}))) \\
 \mathbf{value} &= \text{ReLU}(\text{BatchNorm}(\text{Conv}_v(\mathbf{V}_{\text{sal}}))) \\
 \mathbf{query}_p &= \text{ReLU}(\text{BatchNorm}(\text{Conv}_q(\mathbf{f}_p^{\text{interm}})))
 \end{aligned}
 \tag{3.2.4.4}$$

với Conv_{key} , $\text{Conv}_{\text{value}}$, $\text{Conv}_{\text{query}}$ là các lớp tích chập có kích thước kernel 1×1 , $p = 1, 2, \dots, \frac{H}{2} \times \frac{W}{2}$.

Véc-tơ thể hiện mối quan hệ giữa véc-tơ đặc trưng điểm ảnh với véc-tơ đại diện cho mức độ trọng yếu \mathbf{V}_{sal} :

$$\mathbf{x}_p = \text{sigmoid}(\mathbf{query}_p^T \mathbf{key}) \mathbf{value}
 \tag{3.2.4.5}$$

Véc-tơ đặc trưng dùng để dự đoán bản đồ điểm quan trọng cuối cùng cho từng điểm ảnh có dạng:

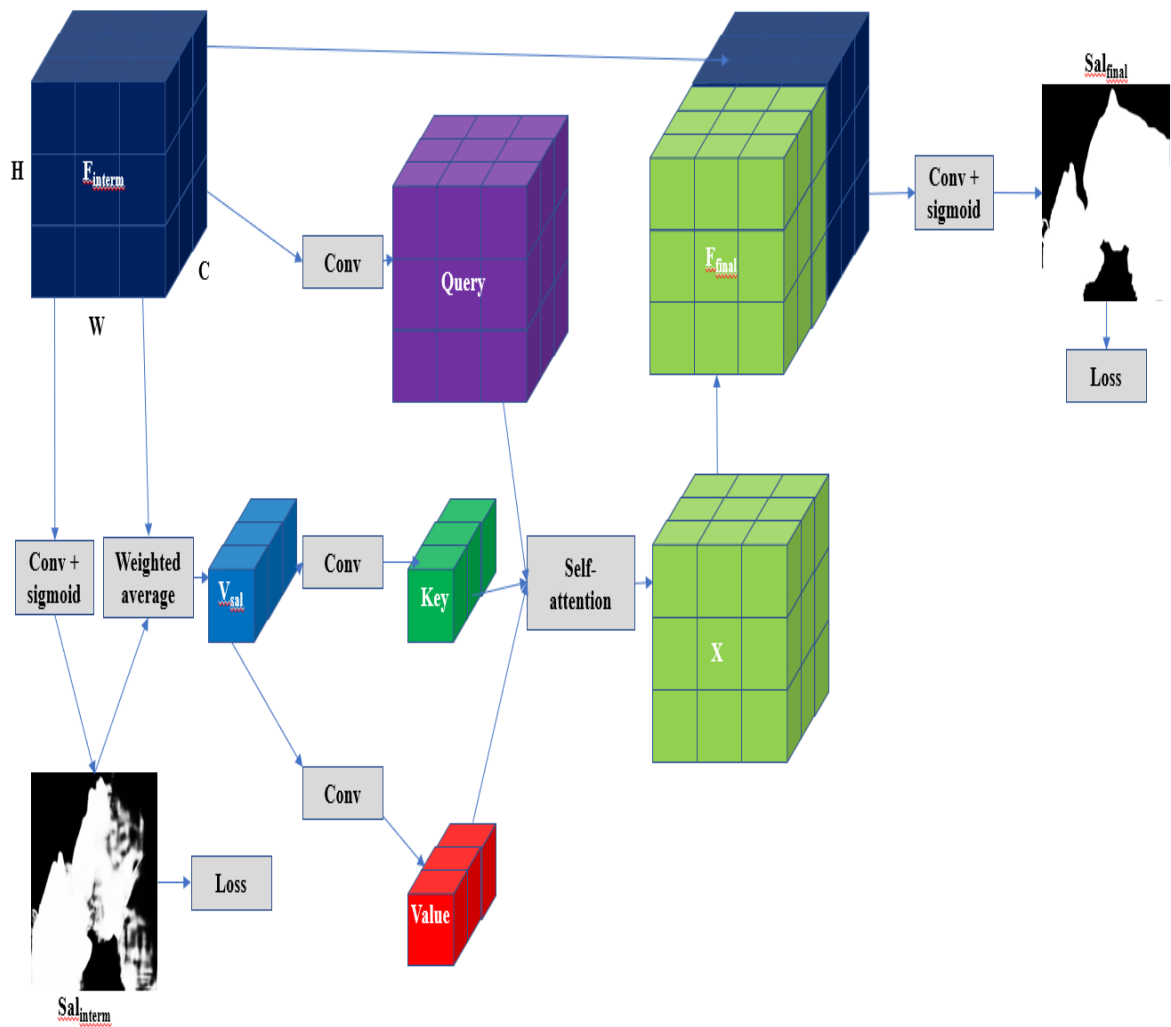
$$\mathbf{f}_p^{final} = \text{concat}(\mathbf{f}_p^{interm}, \mathbf{x}_p), p = 1, 2, \dots, \frac{H}{2} \times \frac{W}{2}. \quad (3.2.4.6)$$

Bản đồ điểm quan trọng cuối cùng được dự đoán:

$$\mathbf{Sal}_{final} = \text{Sigmoid}(\text{Conv}_{final}(\mathbf{F}_{final})) \quad (3.2.4.7)$$

Cơ chế self-attention sẽ điều chỉnh độ trọng yếu của từng véc-tơ đặc trưng điểm ảnh \mathbf{f}_p^{interm} dựa trên mức độ tương đồng giữa nó với véc-tơ đại diện cho mức độ trọng yếu \mathbf{V}_{sal} . Từ đó nhiều sẽ được lược bỏ.

Hình dưới minh họa cơ cách thức hoạt động của **OCR**.



Hình 3.2.4: Cơ chế hoạt động Object-Contextual Representations(OCR) [\[11\]](#)
dựa trên self-attention

3.3 Tính toán hàm lỗi

Hàm lỗi được đề xuất đánh mạnh trọng số vào các điểm ảnh có khả năng dự đoán sai, khó có thể dự đoán chính xác như biên cạnh, đường viền của vật thể (**hình 3.3.1**). Trọng số này được xác định:

$$W_{i,j} = 0.5 \sum_{k \in \{3,15,31\}} \left| \frac{\sum_{h=i-\frac{k}{2}, w=j-\frac{k}{2}}^{h=i+\frac{k}{2}, w=j+\frac{k}{2}} y_{h,w}}{k^2} - y_{i,j} \right| y_{i,j} \quad (3.2.4.1)$$

với y là nhãn, $i = 1, 2, \dots, H, j = 1, 2, \dots, W$.



Hình 3.3.1: Ảnh đầu vào (bên trái), saliency map (giữa), trọng số (bên phải).

Các hàm lỗi được sử dụng là Binary Cross Entropy (BCE) và IoU, L1, kết hợp với trọng số ở trên.

$$\begin{aligned}
\mathbf{BCE}_W &= - \frac{\sum_{i=1}^H \sum_{j=1}^W (1+W_{ij}) \sum_{c=0}^1 (y_c \log(\hat{y}_c) + (1-y_c) \log(1-\hat{y}_c))}{\sum_{i=1}^H \sum_{j=1}^W (1+W_{ij})} \\
\mathbf{IoULoss}_W &= 1 - \frac{\sum_{i=1}^H \sum_{j=1}^W (y_{ij} \hat{y}_{ij})(1+W_{ij})}{\sum_{i=1}^H \sum_{j=1}^W (y_{ij} + \hat{y}_{ij} - y_{ij} \hat{y}_{ij})(1+W_{ij})} \\
\mathbf{L1Loss}_W &= \sum_{i=1}^H \sum_{j=1}^W |y_{ij} - \hat{y}_{ij}| w_{ij}
\end{aligned}
\tag{3.2.4.2}$$

với y , \hat{y} lần lượt là saliency map được dự đoán bởi mô hình và nhãn của nó.

$$\mathbf{Adaptive Intensity Pixel Loss(API)} = \mathbf{BCE}_W + \mathbf{IoULoss}_W + \mathbf{L1Loss}_W
\tag{3.2.4.3}$$

Mô hình phân lớp hai bản đồ điểm quan trọng là **Sal_{interm}** và **Sal_{final}**. Từ hai bản đồ, các hàm mất mát **API** được tính tương ứng **Loss_{interm}** và **Loss_{final}**, kết hợp chúng lại được hàm lỗi cuối cùng, phục vụ cho việc huấn luyện đầu cuối:

$$\mathbf{Loss} = 0.4\mathbf{Loss}_{interm} + \mathbf{Loss}_{final}
\tag{3.2.4.4}$$

3.4 Kết luận

Trong chương này, nhóm đã trình bày chi tiết cũng như đưa ra các hình minh họa cho các giai đoạn chính bao gồm rút trích đặc trưng, chọn lọc, tích hợp đặc trưng và hàm lỗi được sử dụng.

Mô hình được nhóm đề xuất là sự kết hợp giữa PFSNet, Object-Contextual Representations và hàm lỗi Adaptive Intensity Pixel. Cả qui trình có thể được xem là một giai đoạn và được huấn luyện đầu cuối.

Trong chương tiếp theo – chương 4, khóa luận sẽ nói về chi tiết cài đặt, quá trình huấn luyện và kiểm thử. Nhóm cũng sẽ đưa ra các phân tích, đánh giá các kết quả thử nghiệm trên các tập dữ liệu chuẩn. Cuối cùng so sánh kết quả thu được so với mô hình PFSNet không sử dụng OCR và các mô hình hiện đại khác.

Chương 4

Kết quả thử nghiệm

Trong chương này, khóa luận đầu tiên sẽ đề cập đến các bộ dữ liệu chuẩn được dùng để huấn luyện và kiểm thử. Ở mục 4.1 này, công tác chuẩn bị và tăng cường số lượng dữ liệu được nhóm trình bày, từ đó có cái nhìn tổng quan về các bộ dữ liệu.

Sau đó ở mục 4.2, các độ đo được sử dụng để đánh giá mô hình trong giai đoạn kiểm thử được giới thiệu, phân tích ưu nhược điểm của chúng.

Trong mục tiếp theo 4.3, nhóm sẽ đưa ra các kiến trúc cùng siêu tham số của kiến trúc đó mà nhóm muốn khảo sát. Quá trình tinh chỉnh và huấn luyện các mô hình khác nhau trên cũng được đề cập đến.

Nhóm sẽ so sánh các kết quả trong quá trình đánh giá các mô hình với nhau ở mục 4.4.

4.1 Công tác chuẩn bị dữ liệu

4.1.1 Bộ dữ liệu cho quá trình huấn luyện

Nhóm sử dụng bộ dữ liệu **DUTS-TR** cho quá trình huấn luyện và đánh giá độ chính xác của mô hình trong quá trình này. Bộ dữ liệu khá lớn và phổ biến, bao gồm hơn 10553 bức ảnh được chọn lọc và đánh nhãn từ tập ImageNet (chi tiết cách thức đánh nhãn xem mục [1.3.1](#)). Từ năm 2017 đến nay, các mô hình học sâu nghiên cứu về đề tài cùng tên hầu hết sử dụng tập dữ liệu này cho quá trình huấn luyện bởi kích thước lớn cũng như khả năng tổng quát của nó được kiểm chứng trên các tập dữ liệu chuẩn khác (đề cập ở mục sau).

Nhận thấy ở các nghiên cứu trước sử dụng tất cả ảnh trong **DUTS-TR** để huấn luyện mà không có tập validation nên kết quả thu được thiếu tính công bằng, nhóm

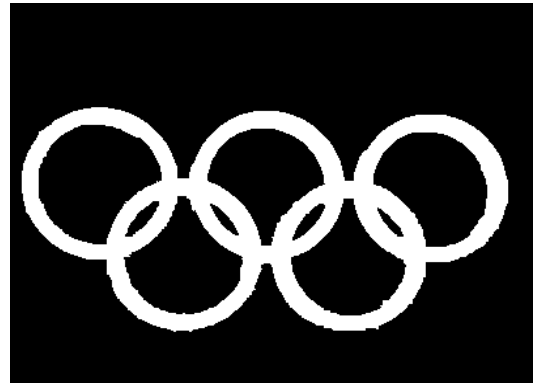
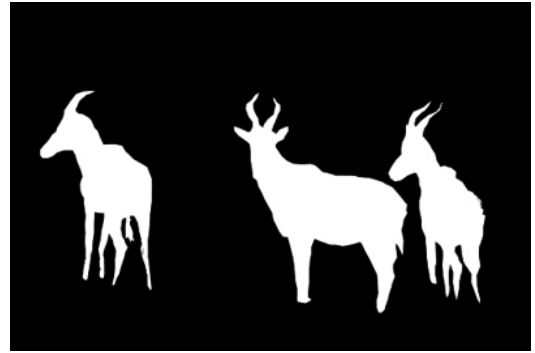
chia bộ dữ liệu ra hai tập train và validation theo tỉ lệ 95:5. Từ việc theo dõi độ lỗi trên tập validation, nhóm sẽ so sánh mô hình với các siêu tham số khác nhau.

Các phương pháp tăng cường dữ liệu cũng được áp dụng để cho mô hình có thể học được nhiều trường hợp thực tế hơn, bao gồm:

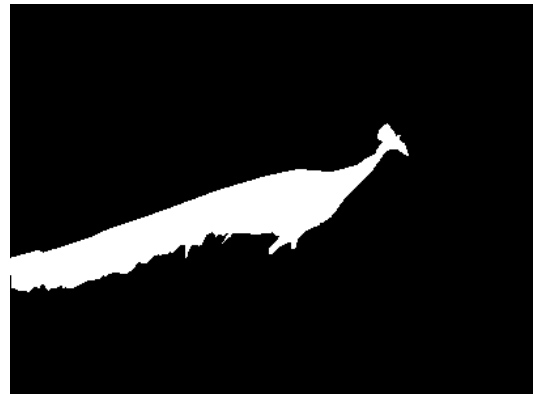
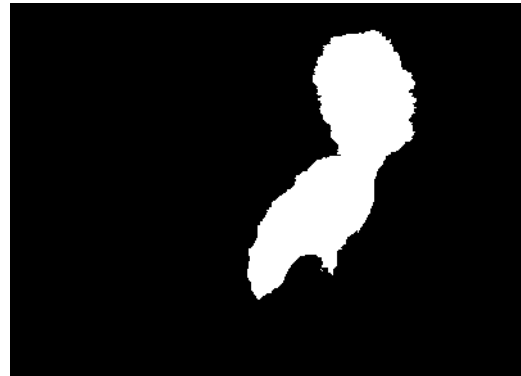
- Lật hình theo chiều ngang.
- Lật hình theo chiều dọc.
- Ngẫu nhiên xoay ảnh 90 độ.
- Làm mờ và nhiễu Gaussian.
- Ngẫu nhiên điều chỉnh độ tương phản, độ sáng của ảnh.

4.1.2 Các bộ dữ liệu cho quá trình kiểm thử

Các bộ dữ liệu **ECSSD**, **DUTS-TE** được sử dụng để kiểm tra khả năng chính xác cũng như tổng quát của mô hình (chi tiết về các bộ dữ liệu ở mục [1.3.1](#)). Các thử thách và khó khăn mà mô hình có thể giải quyết tốt như dự đoán nhiều vật thể trọng yếu trong cùng một bức ảnh hay vật thể đó có màu sắc hòa lẫn vào phông nền, các vật thể có kích thước lớn nhỏ khác nhau, và có thể nằm bất cứ vị trí nào trên bức ảnh. Chi tiết xem **hình 4.1.2.1** và **4.1.2.2**.



Hình 4.1.1: Dữ liệu đánh giá gồm các ảnh có nhiều vật thể trọng yếu



Hình 4.1.2: Vật thể trọng yếu có màu tương đồng với phông nền xung quanh.

4.2 Phương pháp đánh giá

Các độ đo như F-measure, S-measure, MAE được sử dụng để đánh giá khả năng chính xác của mô hình trên các tập dữ liệu kiểm thử **DUTS-TE, ECSSD**.

4.2.1 Độ chính xác – Độ phủ

Độ chính xác tính toán tỷ lệ điểm ảnh dự đoán đúng so với ảnh dự đoán, trong khi đó độ phủ đo đặc tỷ lệ điểm ảnh dự đoán đúng so với ảnh được đánh nhãn.

Gọi bản đồ điểm ảnh quan trọng được dự đoán và đánh nhãn lần lượt là P và G , công thức tính độ chính xác – độ phủ như sau:

$$Precision = \frac{|P \cap G|}{|P|}$$

$$Recall = \frac{|P \cap G|}{|G|}$$

(4.2.1.1)

4.2.2 F-measure

Giữa độ chính xác và độ phủ có sự đánh đổi, nghĩa là độ chính xác cao (có thể chỉ dự đoán một phần đối tượng trọng yếu) thì độ phủ sẽ thấp, và ngược lại nếu độ phủ cao (là vô nghĩa nếu tất cả điểm ảnh được dự đoán là trọng yếu), độ chính xác giảm. Vì hai độ đo này không thể đánh giá toàn diện được mô hình nên nhóm sẽ sử dụng F-measure. F-measure là sự kết hợp độ chính xác và độ phủ thành một độ đo duy nhất, được tính toán như sau:

$$F_{\beta} = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 Precision + Recall}$$

(4.2.2.1)

Với tham số β được gán bằng 0.3 trong hầu hết các bài báo nghiên cứu để đặt trọng số nhiều hơn vào độ chính xác vì như đã đề cập, độ phủ cao quá cũng không có ý nghĩa gì.

4.2.3 MAE – Giá trị độ lỗi trung bình tuyệt đối

Rõ ràng các độ đo trên không tập trung vào những điểm ảnh được dự đoán đúng là không trọng yếu. Các độ đo này chỉ giúp dự đoán mức độ trọng yếu cực kì cao trên vật thể quan trọng nhưng lại không thể dự đoán chính xác các vùng không trọng yếu. Giá trị độ lỗi trung bình tuyệt đối (**MAE**) giải quyết vấn đề này khi sai số của tất cả điểm ảnh có trọng số là như nhau. Giá trị độ lỗi trung bình tuyệt đối được tính toán như sau:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |P(i,j) - G(i,j)|$$

(4.2.3.1)

Tuy nhiên vì xem mọi điểm ảnh là như nhau, MAE thể hiện không tốt với các bức ảnh chứa vật thể nhỏ khi phần lớn điểm ảnh MAE cố gắng dự đoán đúng là điểm không trọng yếu.

4.2.4 S-measure

S-measure đánh giá sự tương đồng về mặt cấu trúc giữa bản đồ điểm quan trọng được dự đoán và nhãn. Độ đo xem xét sự giống nhau về cấu trúc vật thể cũng như cấu trúc vùng. S-measure được tính toán như sau:

$$S = \alpha \times S_o + (1 - \alpha) \times S_r$$

(4.2.4.1)

Với α được gán trong thực tế bằng 0.5.

4.2.5 Kết luận

Với những phân tích trên cũng như sự phổ biến của các độ đo F, S và MAE, nhóm lựa chọn đánh giá mô hình sau khi huấn luyện dựa trên giá trị của chúng.

4.3 Chi tiết cài đặt, quá trình tinh chỉnh và huấn luyện

Nhóm huấn luyện mô hình PFSNet kết hợp OCR với hàm lỗi API (như đã đề cập ở chương 3) với hai cấu hình chính:

- Cấu hình **thứ nhất**: giữ nguyên kiến trúc PFSNet, số lượng kênh của khóa (key channels) trong cơ chế self-attention OCR được gán bằng 64.
- Cấu hình **thứ hai**: giống cấu hình thứ nhất nhưng tăng số kênh khóa lên gấp đôi là 128.

Nhóm sử dụng **NVIDIA GeForce RTX 2070 8Gb**, chạy trên server và sử dụng Pytorch cho quá trình cài đặt.

Nhóm không sử dụng pretrained PFSNet khi huấn luyện mô hình có sử dụng OCR. Chi tiết quá trình tinh chỉnh, huấn luyện, số lượng tham số (PFSNet gốc là 31.180.161) và thời gian tương ứng được trình bày ở **bảng 4.3.1**.

Validation loss được sử dụng để dừng huấn luyện mô hình sớm khi có dấu hiệu độ lỗi này không giảm sau số lần chạy epoch cố định cũng như giảm tốc độ học đi 10 lần khi train loss không giảm.

	PFSNet+OCR (key channels = 64)	PFSNet+OCR (key channels = 128)
Số lượng tham số	31.329.282	31.625.858
Quá trình huấn luyện	<p>Cả mô hình được huấn luyện đầu cuối với tốc độ học (learning rate) cho backbone Resnet là 0.005, các mô-đun còn lại là 0.05. Weight decay cho cả mô hình là 0.0005. Kích thước lô là 8. Kích thước ảnh đầu vào là 352.</p> <p>Chạy hơn 100 epochs với thời gian chạy ~10 phút/epoch.</p> <p>Thời gian cả quá trình là 18-19 giờ.</p>	<p>Huấn luyện với các siêu tham số như cấu hình 1.</p> <p>Chạy hơn 100 epochs với thời gian chạy ~13 phút/epoch.</p> <p>Thời gian cả quá trình là 22-23 tiếng.</p>

Bảng 4.3.1 Chi tiết quá trình tinh chỉnh và huấn luyện trên các cấu hình khác

4.4 Đánh giá và so sánh các cấu hình

Sau quá trình huấn luyện ở mục trước, nhóm đánh giá các mô hình trên các tập dữ liệu kiểm thử ở mục **4.1.2** thu được các kết quả đo độ đo như sau:

Bộ dữ liệu	DUTS-TE			ECSSD		
Độ đo	Avg F	S	MAE	Avg F	S	MAE
Mô hình						
PFS (kết quả theo paper)			0.0360			0.0310
PFS (chạy lại)	0.8533	0.8924	0.0359	0.9260	0.9298	0.0314
PFS OCR 64	0.8390	0.8764	0.0376	0.9228	0.9230	0.0323
PFS OCR 128	0.8538	0.8864	0.0358	0.9280	0.9264	0.0302

Bảng 4.4.1: Kết quả đánh giá các mô hình trên các bộ dữ liệu kiểm thử. 64, 128 lần lượt là số kênh trong mô-đun OCR, theo hai cấu hình nhóm khảo sát.

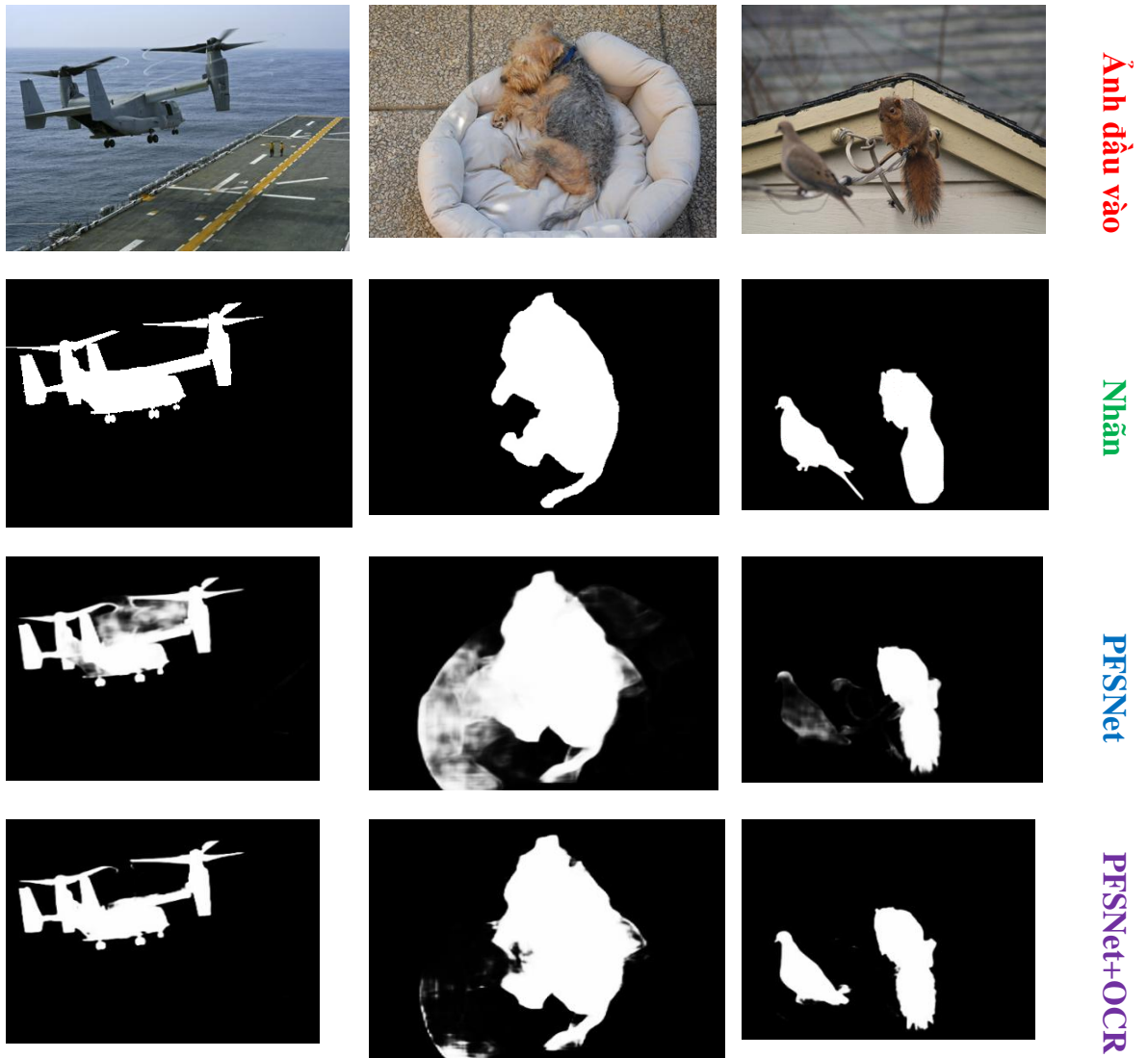
Từ bảng thống kê kết quả trên, nhóm đưa ra các nhận xét sau đây:

- Nhìn chung, mô hình thể hiện tốt trên tập dữ liệu này cũng cho kết quả tốt trên tập dữ liệu còn lại. Từ đó cũng cho thấy tính tổng quát của tập dữ liệu huấn luyện.
- Kết quả trên tập dữ liệu ECSSD và DUTS-TE cho thấy độ khó của DUTS-TE cao hơn.
- Mô hình cải tiến cho kết quả sát sao với mô hình gốc. PFSNet sử dụng OCR với số kênh gấp đôi của khối đặc trưng đầu vào (128 kênh) nhỉnh hơn một chút.

Tuy nhiên, những con số sát sao này chưa cho thấy sự hiệu quả của mô-đun đề xuất OCR cũng như mục đích được kì vọng ban đầu mà mô-đun đem lại. Nhóm sẽ khảo sát đầu ra (bản đồ điểm quan trọng) của PFSNet gốc và PFSNet có sử dụng OCR cho kết quả tốt nhất (128 kênh). Từ đó sẽ chỉ ra OCR hoạt động tốt đúng như mong đợi của nhóm ban đầu.

Như đã đề cập ở các phần trước, hai mô hình gốc và cải tiến phân lớp trên khối đặc trưng đã được tích hợp thông qua một lớp tích chập theo sau bởi hàm sigmoid và không áp dụng bất kì kĩ thuật hậu xử lý nào.

Hình 4.4.1 cho thấy OCR có thể đồng thời loại bỏ các vùng, điểm ảnh không trọng yếu (như nệm của chú chó và vùng biển giữa các cánh máy bay) và khôi phục các vật thể quan trọng (chú chim bồ câu).

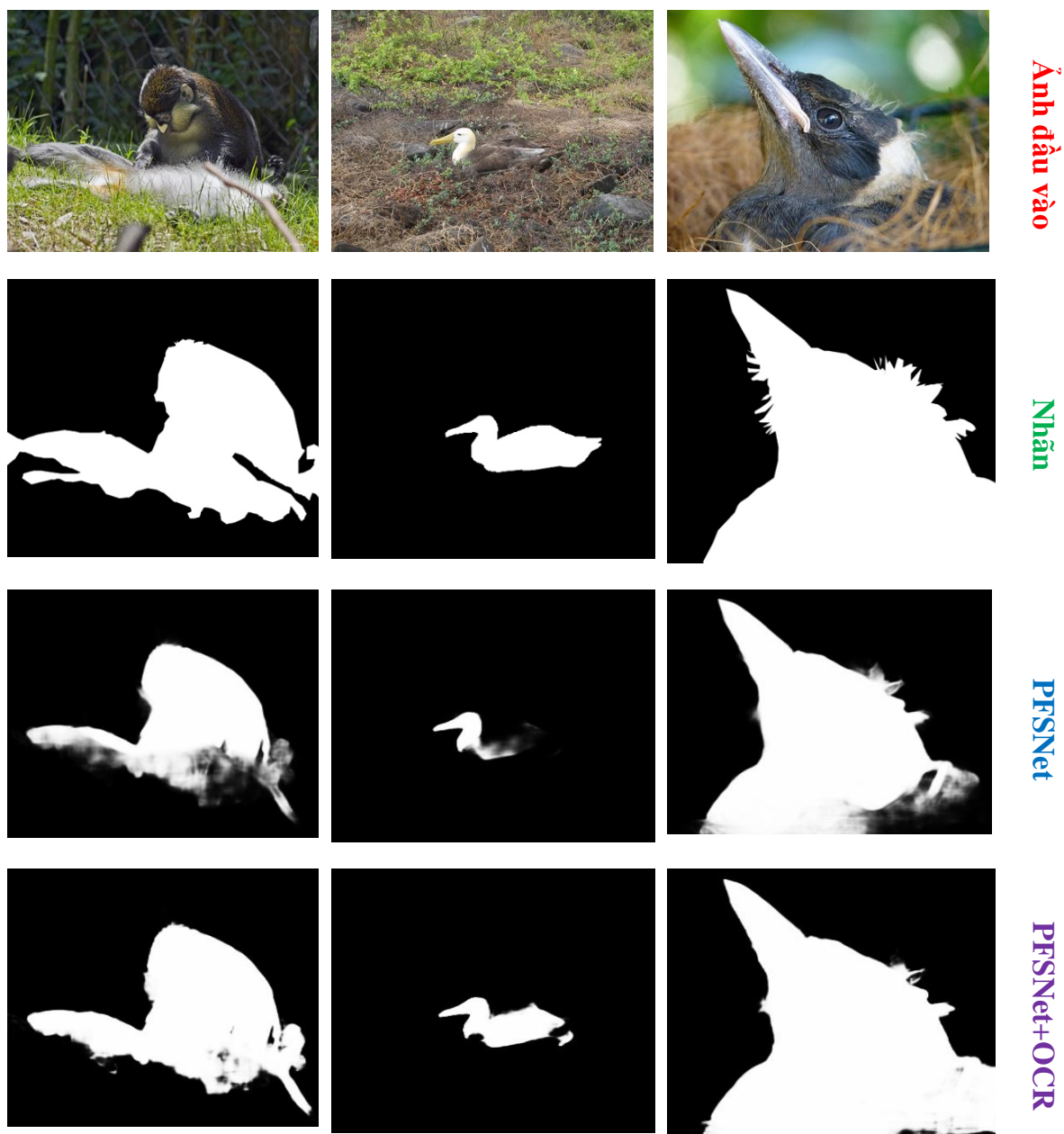


Hình 4.4.1: Mô hình cải tiến cho loại bỏ các vùng, điểm ảnh không trọng yếu, trong khi giữ lại và tăng cường các đối tượng trọng yếu.

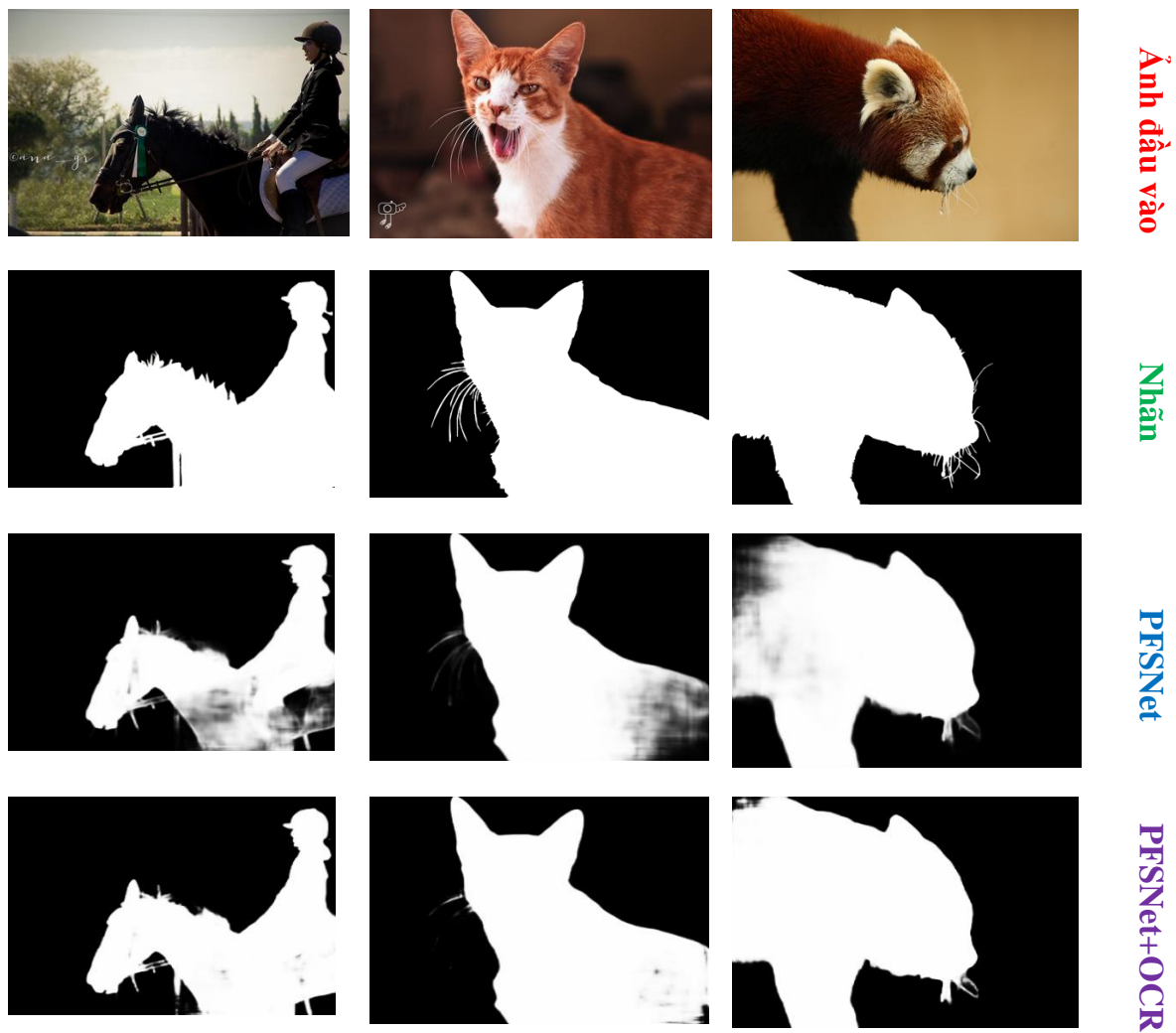
Với các bức ảnh thử thách, khó khăn hơn, mô hình cải tiến cũng cho thấy kết quả tốt hơn so với mô hình gốc:

- Các bức ảnh cùng màu hoặc nằm sau phong nền, PFSNet sử dụng OCR vẫn có thể phát hiện được (xem **hình 4.4.2**).
- Các vật thể chiếm phần lớn bức ảnh (**hình 4.4.3**) hay các vật thể nhỏ, mỏng (**hình 4.4.4** và **4.4.5**) được mô hình tăng cường giá trị trọng yếu, hạn chế bỏ sót các vật thể đó, cũng như loại bỏ đi nhiễu (bức tường tuyết trong **hình 4.4.4** và đám mây hoặc bóng của cột tuabin gió trong **hình 4.4.5**).

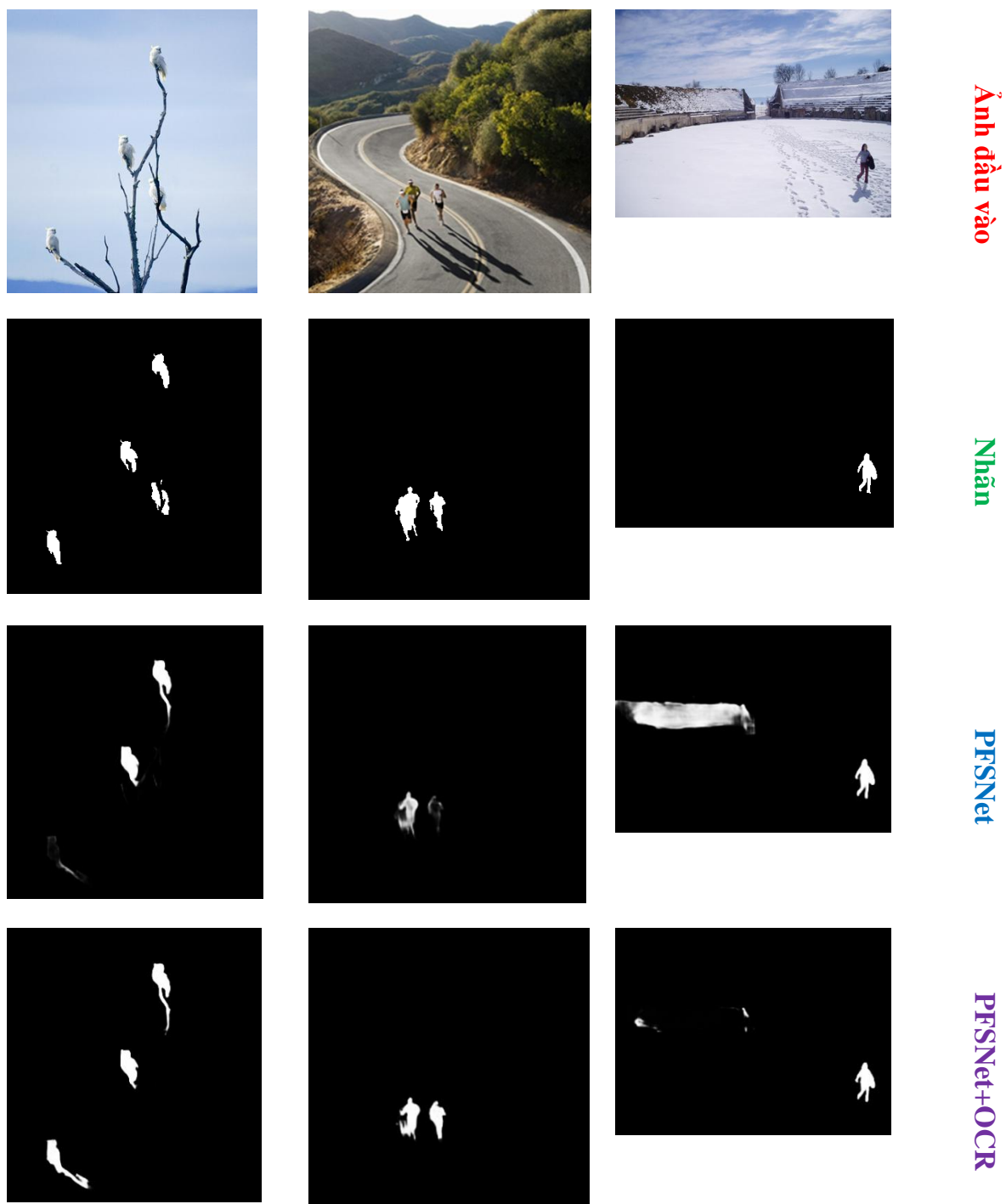
Mô-đun đề xuất cho kết quả giống như mục đích kì vọng của chúng, có thể linh hoạt loại bỏ các vật thể không trọng yếu cũng như khôi phục đối tượng quan trọng. Sự trọng yếu của véc-tơ đặc trưng của từng điểm ảnh được điều chỉnh thích hợp dựa trên mối tương đồng với véc-tơ đại diện cho sự trọng yếu (như đã bàn luận ở trên). Các đối tượng được phát hiện chắc chắn có mức độ trọng yếu cao thể hiện qua độ sáng trong ảnh dự đoán.



Hình 4.4.2: Mô hình cải tiến cho kết quả phân lớp tốt hơn nhiều so với PFSNet trên các bức ảnh bị che hoặc cùng màu với phong nền (occlusion).



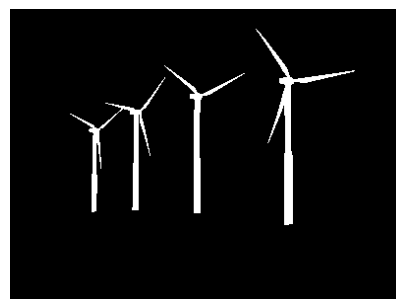
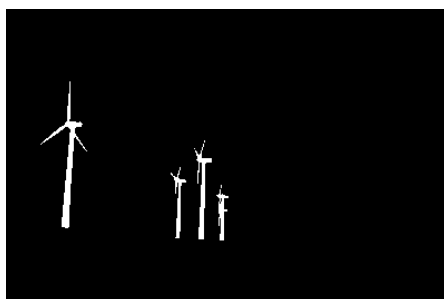
Hình 4.4.3: Kết quả trên các bức ảnh chứa vật thể trọng yếu không toàn vẹn, vị trí sát biên ảnh, lớn hơn nửa bức ảnh.



Hình 4.4.4: Kết quả trên các bức ảnh chứa một hoặc nhiều vật thể có kích thước nhỏ.



Ảnh đầu vào



Nhân



PFSNet



PFSNet+OCR

Hình 4.4.5: Kết quả trên các bức ảnh chứa nhiều vật thể có hình dáng mỏng.

4.5 Kết luận

Trong chương này, nhóm đã trình bày cả qui trình chuẩn bị dữ liệu, xây dựng mô hình, huấn luyện, đánh giá và so sánh các mô hình đề xuất. Mặc dù kết quả số liệu không cho thấy sự cách biệt (một phần là do mô hình PFSNet đã rất tốt, nên đề cải thiện bức phá là thử thách lớn), nhưng khi khảo sát các kết quả đầu ra, nhóm đã cho thấy khả năng mà OCR có thể cung cấp để cải thiện mô hình gốc.

Từ những phân tích trên, nhóm sẽ đưa ra các hướng phát triển cũng như thử nghiệm hứa hẹn có thể cải thiện kết quả nhiều hơn trong chương cuối cùng.

Vậy là ta đã đi qua các nội dung mà khóa luận này muốn truyền tải. Trong chương cuối cùng, nhóm sẽ tổng kết các đóng góp của khóa luận, cũng như đưa ra các hướng đi trong tương lai.

Chương 5

Tổng kết và hướng phát triển

5.1 Tổng kết

Phát hiện đối tượng trọng yếu trong ảnh là một bài toán khó vì nó liên quan đến nhiều lĩnh vực và cơ chế tập trung thị giác vẫn chưa được giải đáp. Đề tài có ý nghĩa vô cùng to lớn cả trong học thuật khi cố gắng mô phỏng khả năng nhận thức thị giác của con người, lẫn ứng dụng thực tế trong các tác vụ yêu cầu rút trích hoặc tóm tắt thông tin quan trọng trong bức ảnh.

Từ những phân tích các nghiên cứu liên quan và những thách thức mà bài toán đang gặp phải, khóa luận đã chỉ ra việc sử dụng mạng học sâu thực sự phù hợp khi mô hình học gián tiếp cơ chế tập trung thị giác thông qua bộ dữ liệu được đánh nhãn bởi chính con người có khả năng nhận thức bình thường.

Bài toán dự đoán chính xác trên từng điểm ảnh nên quá trình rút trích và tích hợp các khối đặc trưng đa bậc hiệu quả ảnh hưởng rất lớn đến kết quả của mô hình. Dựa trên điều này nhóm đã lựa chọn khảo sát và cải tiến trên mô hình gốc PFSNet. Quá trình rút trích, chọn lọc và tích hợp đặc trưng của mô hình này chỉ thực hiện trên hai khối đặc trưng liền kề, có ít sự cách biệt nhất về kích thước cũng như mức độ thông tin.

Việc có thể cung cấp thông tin toàn cục cho từng điểm ảnh cực kì quan trọng vì để xác định (các) đối tượng nào là trọng yếu, mô hình phải nhìn được toàn bộ bức ảnh. Từ đó dẫn đến sự khác nhau trong quá trình thiết kế mô hình của bài toán khác với các đề tài như phân đoạn ngữ nghĩa (khi mà mỗi điểm ảnh chỉ cần thông tin của những vùng ảnh lân cận).

Nhóm đã cải tiến mô hình gốc bằng cách áp dụng mô-đun Object Contextual Representation (OCR) trong bài toán phân đoạn ngữ nghĩa. Kết quả thử nghiệm cho thấy OCR hoạt động như kì vọng khi điều chỉnh mức độ trọng yếu của từng điểm

ảnh dựa trên mức độ trọng yếu của tất cả các điểm ảnh khác. Tuy nhiên mô-đun chỉ hoạt động tốt khi các véc-tơ đặc trưng của các điểm ảnh tương ứng có thể độc lập phân lớp. Từ đó có thể tự động lược bỏ đối tượng không trọng yếu và khôi phục vật thể quan trọng.

Qua những phân tích trên, bất kì kiến trúc, kĩ thuật nào từ các đề tài khác có thể rút trích và tích hợp các đặc trưng đa bậc giàu thông tin toàn cục hiệu quả đều được kì vọng áp dụng tốt cho bài toán phát hiện đối tượng trọng yếu.

5.2 Các hướng nghiên cứu trong tương lai

Như đã phân tích ở chương trước, OCR hoạt động tốt khi nó điều chỉnh mức độ trọng yếu của từng điểm ảnh dựa vào sự giống nhau giữa các véc-tơ đặc trưng điểm ảnh và véc-tơ đại diện cho mức độ trọng yếu trong không gian \mathbf{R}^{128} . Từ đó đưa ra câu hỏi “Nếu ta tăng chiều không gian véc-tơ lên nữa thì liệu mô hình có thể tốt hơn nữa không?”. Theo như OCR được sử dụng thành công trong bài toán phân đoạn ảnh với số kênh mặc định là 256, nhóm khá tự tin mô hình sẽ có thể cải thiện hơn nữa. Khi không gian véc-tơ bây giờ có thể rộng hơn, phân bố của các véc-tơ đặc trưng điểm ảnh sẽ rộng hơn, từ đó mối tương quan giữa chúng với véc-tơ đại diện cho mức độ trọng yếu sẽ rõ ràng hơn.

Tăng số kênh cũng là tăng số lượng tham số cho mô hình, mô hình có thể bắt trọn nhiều tình huống hơn, từ đó cải thiện kết quả. Tuy nhiên, nếu chú ý ta sẽ nhận thấy dấu hiệu bottleneck giữa khối đặc trưng được tích hợp trước khi cho vào OCR. Khối đặc trưng đầu vào của OCR nhận từ PFSNet chỉ có số kênh là 64. Trong một số bài báo như InceptionNetv3, EfficientNet, hay MobileNetv2, bottleneck này có thể sẽ không gây hại. Vì vậy nhóm đưa ra các hướng thử nghiệm nhóm mong muốn thực hiện trong tương lai như sau:

- Đầu tiên, mô hình sẽ được tăng số kênh OCR lên 256. Nếu cho thấy cải thiện sẽ tăng tiếp đến khi không nhận thấy độ chính xác đạt được đáng giá với kích thước mô hình.

- Tăng số kênh trong các mô-đun SEM, AFMs để tránh hiện tượng bottleneck gây ra. Việc tích hợp đặc trưng có thể tốt hơn từ đây. Mặc dù điều này tăng tham số mô hình nhưng so với các mô hình tốt nhất hiện này, điều đó là đáng để thực nghiệm để đạt được độ chính xác cao hơn.
- Nhận thấy OCR có thể hoạt động tốt trên bất kì kiến trúc rút trích và tích hợp đặc trưng đa bậc hiệu quả (như trong bài toán phân đoạn và phát hiện vật thể trọng yếu), nhóm nghĩ dần thay thế PFSNet thành các kiến trúc nhẹ hơn, hiệu quả hơn, có backbone như EfficientNet chẳng hạn.
- Hiện tại, các kiến trúc Transformer đã cho thấy sự vượt trội trong rất nhiều tác vụ (và có thể trong đề tài này trong tương lai). Hướng nghiên cứu này cũng rất đáng để tìm hiểu.

Đó là tất cả những hướng cải tiến mà khóa luận muốn thực hiện trong tương lai. Vì giới hạn thời gian nghiên cứu, thử nghiệm cũng như tài nguyên không cho phép khảo sát, cải tiến mô hình phức tạp nên nhóm đành dừng lại tại đây trong khóa luận này.

Tài liệu tham khảo

- [1] P. Zhang, D. Wang, H. Lu, H. Wang and X. Ruan, "Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 202-211.
- [2] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan and M. Jagersand, "BASNet: Boundary-Aware Salient Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 7471-7481.
- [3] L. Zhang, J. Dai, H. Lu, Y. He and G. Wang, "A Bi-Directional Message Passing Model for Salient Object Detection," 2018 IEEE/CVF Conference on Computer Vision .
- [4] Z. Wu, L. Su and Q. Huang, "Cascaded Partial Decoder for Fast and Accurate Salient Object Detection," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3902-3911.
- [5] G. Li and Y. Yu, "Deep Contrast Learning for Salient Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 478-487.
- [6] N. Liu and J. Han, "DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 678-686.
- [7] Q. Hou, M. -M. Cheng, X. Hu, A. Borji, Z. Tu and P. H. S. Torr, "Deeply Supervised Salient Object Detection with Short Connections," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4, pp. 815-828, 1 April 2019.

- [8] G. Lee, Y. -W. Tai and J. Kim, "Deep Saliency with Encoded Low Level Distance Map and High Level Features," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 660-668.
- [9] L. Wang, H. Lu, X. Ruan and M. -H. Yang, "Deep networks for saliency detection via local estimation and global search," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3183-3192.
- [10] R. Zhao, W. Ouyang, H. Li and X. Wang, "Saliency detection by multi-context deep learning," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1265-1274.
- [11] Yuan, Y., Chen, X., Wang, J. (2020). Object-Contextual Representations for Semantic Segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds) Computer Vision – ECCV 2020. ECCV 2020.
- [12] Ma, M., Xia, C., & Li, J. (2021). Pyramidal Feature Shrinking for Salient Object Detection. Proceedings of the AAAI Conference on Artificial Intelligence, 35(3), 2311-2318.
- [13] N. Liu, J. Han and M. Yang, "PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089-3098, doi: 10.1109/CVPR.2018.00326.
- [14] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling and R. Yang, "Salient Object Detection in the Deep Learning Era: An In-Depth Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3239-3259, 1 June 2022.
- [15] N. Liu, N. Zhang, K. Wan, L. Shao and J. Han, "Visual Saliency Transformer," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4702-4712.

- [16] L.M Seok, S.W Seok, H.S Won, "TRACER: Extreme Attention Guided Salient Object Tracing Network".
- [17] C. Yang, L. Zhang, H. Lu, X. Ruan and M. Yang, "Saliency Detection via Graph-Based Manifold Ranking," 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166-3173.
- [18] L. Wang et al., "Learning to Detect Salient Objects with Image-Level Supervision," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3796-3805.
- [19] Q. Yan, L. Xu, J. Shi and J. Jia, "Hierarchical Saliency Detection," 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 1155-1162.

Phụ lục

Môi trường phát triển

Môi trường phát triển

- Python 3.9.12
- Intel(R) Core (TM) i5-7300HQ CPU @ 2.50GHz 2.50 GHz, Ram 8Gb, GPU RTX 2070 8Gb.

Các thư viện sử dụng

- albumentations 1.2.0
- torchvision 0.12.0
- torch 1.11.0
- scikit-learn 1.0.2
- numpy 1.21.5
- opencv-contrib-python 4.5.5.64
- tqdm 4.64.0