

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

ĐỒ ÁN TỐT NGHIỆP

**Phân vùng polyp trên ảnh nội soi bằng phương pháp
học có giám sát dựa trên kiến trúc Transformer**

MAI VĂN HOÀ

hoa.mv173122@sis.hust.edu.vn

Ngành: Khoa học máy tính

Giảng viên hướng dẫn: TS. Nguyễn Thị Oanh

Chữ ký GVHD

Khoa: Khoa học máy tính

Trường: Công nghệ Thông tin và Truyền thông

HÀ NỘI, 08 / 2022

LỜI CẢM ƠN

*"Nếu em về qua phố nhỏ Bách Khoa
Nhớ gửi cho tôi chùm bằng lăng cuối hạ
Nơi bè bạn ngày đêm hối hả
Mùa thi cuối cùng, mai hết tuổi sinh viên."*

Chiều mùa thu tháng 8, rǎo bước chân vội vã qua những con đường mang tên Bách Khoa để tránh những cơn mưa bất chợt. Bỗng nhận ra quãng thời gian còn được làm sinh viên của mình không còn dài nữa, “sắp đến lúc phải ra đi và bỏ lại những gì đã trải qua”. Vậy là thoảng chốc đã 5 năm đã trôi qua, 5 năm được sống, được học tập, được là một sinh viên Bách Khoa.

Cảm ơn đại gia đình đã luôn động viên, tin tưởng, là chỗ dựa tinh thần của con trong suốt quá trình học tập.

Để viết lên được những dòng này, em xin gửi lời cảm ơn chân thành đến cô Nguyễn Thị Oanh, giảng viên khoa Khoa học máy tính – Trường Công nghệ Thông tin và Truyền thông đã luôn tận tình, quan tâm, hướng dẫn em hoàn thành đồ án tốt nghiệp của mình.

Em cũng xin gửi lời cảm ơn đến toàn bộ thầy cô của Đại học Bách Khoa Hà Nội nói chung và Trường Công nghệ Thông tin và Truyền thông nói riêng đã tạo điều kiện cơ sở vật chất, truyền đạt nhiều kiến thức và kinh nghiệm vô giá tạo nền tảng cho cuộc sống của em sau này.

Cảm ơn anh Thịnh, anh Tâm, anh Công, anh Cam và anh Viên trong công ty cổ phần ThinkLABs đã tạo điều kiện thời gian, quan tâm và động viên em hoàn thành đồ án tốt nghiệp.

Cuối cùng, xin gửi lời cảm ơn đến những người bạn đã cùng tôi trải qua những ngày tháng sinh viên khó khăn, cảm ơn Shin đã dành cả thanh xuân để đồng hành, chia sẻ cùng mình trong suốt quãng thời gian qua.

Cảm ơn Bách Khoa! Cảm ơn vì tất cả những gì đã qua.

TÓM TẮT NỘI DUNG ĐỒ ÁN

Ung thư đại trực tràng (CRC) là bệnh ác tính thường gặp thứ ba ở nam và thứ hai ở nữ, và là nguyên nhân phổ biến thứ ba gây tử vong do ung thư. CRC có thể được ngăn ngừa bằng cách theo dõi nội soi thường quy và cắt bỏ polyp. Tuy nhiên, quá trình nội soi phát hiện polyp phụ thuộc vào nhiều yếu tố như chất lượng hệ thống trang thiết bị, quy trình nội soi, môi trường làm việc và kinh nghiệm của bác sĩ. Do đó, việc áp dụng công nghệ mới như các kỹ thuật nội soi tăng cường hình ảnh, dây soi với góc mở vi trường rộng hay ứng dụng công nghệ thông tin đặc biệt là xây dựng các thuật toán trí tuệ nhân tạo được kỳ vọng hỗ trợ bác sĩ trong chẩn đoán và điều trị ung thư đại tràng, cải thiện độ chính xác, giảm thiểu việc bỏ sót polyp trong quá trình nội soi.

Đồ án hướng tới xây dựng mô hình trí tuệ nhân tạo sử dụng kiến trúc mạng học sâu cho bài toán phân vùng polyp trên ảnh nội soi đại tràng. Để xây dựng một mô hình mạng hiệu quả, trong đồ án tập trung tìm hiểu kiến trúc và cách tích hợp 3 mô-đun: (i) **Mô-đun mã hóa** sử dụng Mix Transformer (MiT) để trích xuất đặc trưng của ảnh đầu vào. Đây là kiến trúc mạng hiệu quả, đang ngày càng được sử dụng phổ biến để thay thế kiến trúc mạng nơron tích chập truyền thống bởi khả năng trích xuất đặc trưng toàn cục và xử lý vấn đề phụ thuộc xa thông qua cơ chế chú ý (self-attention). (ii) **Mô-đun giải mã** tích hợp thông tin từ các đặc trưng cho ra bởi mô-đun mã hóa với sự trợ giúp của cơ chế chú ý theo cả chiều sâu và chiều không gian nhằm tạo ra bản đồ đặc trưng toàn cục, bản đồ đặc trưng này có khả năng xấp xỉ được vị trí tương đối và hình dạng của polyp. (iii) **Mô-đun tinh chỉnh đặc trưng** sử dụng cơ chế chú ý ngược kết hợp với 4 đặc trưng cho ra bởi kiến trúc MiT tại 4 giai đoạn của mạng nhằm tìm ra các chi tiết còn thiếu, tinh chỉnh vùng biên của polyp, bổ sung thông tin cho bản đồ đặc trưng toàn cục nhằm tạo ra kết quả dự đoán cuối cùng đáng tin cậy hơn.

Thực nghiệm cho thấy mô hình đề xuất đạt kết quả tốt trên cả khả năng học và khả năng tổng quát hóa. Đồng thời, số lượng tham số và độ phức tạp tính toán đều nhỏ hơn các phương pháp phân vùng polyp dựa trên Transformer hiện tại.

MỤC LỤC

CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI.....	1
1.1 Bài toán phân vùng polyp đại tràng	1
1.2 Các giải pháp hiện tại và hạn chế.....	3
1.3 Mục tiêu và định hướng giải pháp.....	3
1.4 Đóng góp của đồ án	4
1.5 Bố cục đồ án	4
CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT	5
2.1 Tổng quan bài toán phân vùng ảnh	5
2.2 Mạng nơron tích chập.....	7
2.2.1 Tổng quan mạng nơron tích chập	7
2.2.2 Một số phương pháp tính tích chập	9
2.3 Transformer truyền thống	12
2.4 Vision Transformer (ViT)	14
2.5 Một số nghiên cứu liên quan đến bài toán phân vùng polyp	15
CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT.....	18
3.1 Tổng quan kiến trúc mô hình đề xuất	18
3.2 Mô-đun mã hóa	18
3.3 Mô-đun giải mã	23
3.3.1 Mô-đun giải mã đề xuất	23
3.3.2 Convolutional Block Attention Module (CBAM)	25
3.4 Mô-đun tinh chỉnh đặc trưng.....	28
3.4.1 Trích xuất đặc trưng kim tự tháp theo kênh (CFP)	28
3.4.2 Mô-đun chú ý ngược (RA) và kết nối phần dư.....	29
3.5 Định nghĩa hàm măt măt	31

CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM.....	33
4.1 Độ đo đánh giá	33
4.1.1 Intersection over Union (IoU, Jaccard index)	33
4.1.2 Dice coefficient	34
4.2 Dữ liệu	35
4.3 Phương pháp tiến hành thực nghiệm	36
4.4 So sánh kết quả thực nghiệm các mô hình	38
4.5 Đánh giá ảnh hưởng của các mô-đun	41
4.6 Đánh giá ảnh hưởng của các phương pháp khởi tạo trọng số	43
CHƯƠNG 5. KẾT LUẬN	45
5.1 Kết luận	45
5.2 Hướng phát triển trong tương lai	45
TÀI LIỆU THAM KHẢO.....	49

DANH MỤC HÌNH VẼ

Hình 1.1:	Polyp đại tràng.....	1
Hình 1.2:	Một ví dụ về: (a) ảnh nội soi đại tràng và (b) polyp được đánh nhãn bởi bác sỹ.....	2
Hình 2.1:	Minh họa sự khác biệt giữa bài toán phân vùng ngữ nghĩa và phân vùng cá thể	5
Hình 2.2:	Kiến trúc cơ bản của một mạng nơron tích chập.....	8
Hình 2.3:	Minh họa tính tích chập hai chiều.....	9
Hình 2.4:	Minh họa tính tích chập 3 chiều	10
Hình 2.5:	(a) Phép tích chập thông thường, (b) Phép tích chập dãn nở với tỉ lệ dãn nở bằng hai.	11
Hình 2.6:	(a) Phép tích chập thông thường, (b) Phép tích chập tách biệt chiều sâu.	11
Hình 2.7:	Kiến trúc Transformer truyền thống.....	12
Hình 2.8:	Kiến trúc Vision Transformer	14
Hình 2.9:	Kiến trúc mạng U-Net	16
Hình 3.1:	Tổng quan kiến trúc đề xuất.....	18
Hình 3.2:	Kiến trúc Mix Transformer.....	20
Hình 3.3:	Overlapping patch embedding.	20
Hình 3.4:	Mô-đun Transformer trong MiT.....	21
Hình 3.5:	Cơ chế tự chú ý trong: (a) Transformer truyền thống, (b) MiT.....	21
Hình 3.6:	Mix-FFN.....	22
Hình 3.7:	Các phiên bản của kiến trúc Mix Transformer	23
Hình 3.8:	Mô-đun giải mã trong bài báo SegFormer.....	24
Hình 3.9:	Mô-đun giải mã đề xuất.	25
Hình 3.10:	Tổng quan kiến trúc CBAM.	25
Hình 3.11:	Minh họa cách tích hợp CBAM vào kiến trúc cơ sở.	26
Hình 3.12:	Mô-đun chú ý theo kênh.	26
Hình 3.13:	Mô-đun chú ý theo không gian	27
Hình 3.14:	Tổng quan mô-đun tinh chỉnh đặc trưng.	28
Hình 3.15:	(a) Mô-đun CFP, (b) Mô-đun FP.	29
Hình 3.16:	Mô-đun chú ý ngược (RA).	30
Hình 3.17:	Kết nối phần dư bổ sung thông tin trích xuất được từ mô-đun RA cho đầu ra dự đoán của bước trước S_{i-1}	30

Hình 4.1:	Minh họa hình học độ đo IoU.....	33
Hình 4.2:	Minh họa hình học độ đo Dice.....	34
Hình 4.3:	(a) Ma trận nhầm lẫn và (b) Minh họa hình học của ma trận nhầm lẫn trong trường hợp phân vùng có hai nhãn lớp.....	35
Hình 4.4:	So sánh độ chính xác các phương pháp khác nhau.	39
Hình 4.5:	Một số kết quả dự đoán của mô hình đề xuất và các mô hình khác.	40
Hình 4.6:	Một số nhược điểm của mô hình đề xuất.	40
Hình 4.7:	Ảnh hưởng của các mô-đun đến độ chính xác của mô hình... .	42
Hình 4.8:	Kết quả dự đoán tại các đầu ra biên của mô hình đề xuất....	43

DANH MỤC BẢNG BIỂU

Bảng 4.1:	Thống kê thông số các bộ dữ liệu	36
Bảng 4.2:	Thống kê thông số phần cứng trên Google Colab	36
Bảng 4.3:	Thống kê thông số của mô hình đề xuất	36
Bảng 4.4:	So sánh hiệu năng của các phương pháp trên 5 tập kiểm tra: Kvasir, ClinicDB, ColonDB, EndoScene (CVC-T) và ETIS..	38
Bảng 4.5:	Số lượng tham số và độ phức tạp của các phương pháp khác nhau.....	39
Bảng 4.6:	Ảnh hưởng của các mô-đun đến khả năng học và khả năng tổng quát hóa của mô hình.....	42
Bảng 4.7:	Ảnh hưởng của các phương pháp khởi tạo trọng số	43

DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT

Thuật ngữ	Ý nghĩa
CAM	Mô-đun chú ý theo kênh (Channel attention module)
CFP	Trích xuất đặc trưng kim tự tháp theo kênh (Channel-wise Feature Pyramid)
CNN	Mạng nơron tích chập (Convolutional Neural Network)
CRC	Ung thư đại trực tràng (Colorectal cancer)
RA	Chú ý ngược (Reverse Attention)
SAM	Mô-đun chú ý theo không gian (Spatial attention module)
SOTA	State-of-the-art

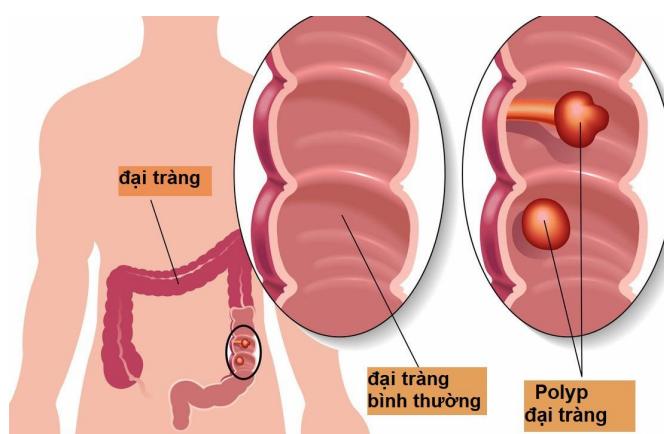
CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

Trong chương này sẽ trình bày ngũ cảnh của bài toán, tổng quan một số giải pháp hiện tại và các thách thức còn tồn tại. Từ đó đặt ra mục tiêu và định hướng giải pháp trong đồ án để giải quyết bài toán.

1.1 Bài toán phân vùng polyp đại tràng

Hiện nay Việt Nam là một quốc gia có tỷ lệ dân số khá đông khoảng hơn 98 triệu dân, gánh nặng bệnh tật tương đối lớn đặc biệt là các bệnh lý về tiêu hóa, gan mật. Tuy nhiên, khả năng đáp ứng của các cơ sở y tế còn rất nhiều hạn chế về kỹ thuật cũng như đội ngũ y tế. Theo ước tính, số lượng bác sĩ nội soi tiêu hóa chỉ đáp ứng được nhu cầu của 5-10% dân số¹. Điều này đặt ra những thách thức to lớn đối với việc chẩn đoán chính xác và điều trị bệnh.

Ung thư đại tràng (CRC) là nguyên nhân phổ biến thứ ba gây tử vong liên quan đến ung thư trên thế giới, chiếm 5.8% tổng số ca tử vong do ung thư vào năm 2018. Ung thư đại tràng thường bắt đầu lành tính (gọi là polyp). Polyp không phải là u nhưng là một tổn thương có hình dạng giống như một khối u, có cuống hoặc không, do niêm mạc đại tràng và tổ chức dưới niêm mạc tăng sinh tạo thành. Polyp không phải là ung thư và hầu hết là vô hại nhưng chúng có thể phát triển thành ung thư sau một thời gian dài, gây tử vong khi được tìm thấy ở giai đoạn muộn của nó. Hình 1.1 minh họa đại tràng bình thường và đại tràng có chứa polyp.



Hình 1.1: Polyp đại tràng².

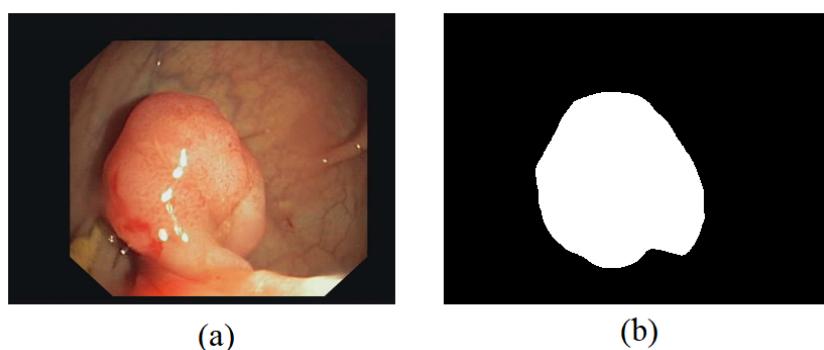
Ngăn ngừa CRC bằng tầm soát polyp thường xuyên, xét nghiệm sàng lọc và loại bỏ các tổn thương u tuyến đại tràng là rất quan trọng và đã trở thành một trong

¹<https://nhandan.vn/viet-nam-som-ung-dung-tri-tue-nhan-tao-trong-noi-soi-phat-hien-polyp-dai-trang-post621914.html>

²https://www.vinmec.com/vi/ung-buou-xa-tri/thong-tin-suc-khoe/tri-tue-nhan-tao-ai-co-giup-danh-gia-dac-diem-polyp-dai-trang-nhu-nao/?link_type=related_posts

những ưu tiên sức khỏe trên toàn thế giới. Nội soi đại tràng là một phương pháp tầm soát phổ biến, hiệu quả để phát hiện polyp bởi nó có thể đồng thời cung cấp thông tin về vị trí và hình dạng của polyp, từ đó tiến hành sàng lọc, phát hiện và loại bỏ sớm trước khi chúng phát triển thành CRC. Một nghiên cứu đã theo dõi hơn 314 nghìn ca nội soi đại tràng gần đây cho thấy, trong 10 nam giới thì tỷ lệ phát hiện u tuyến trong đại tràng tăng mỗi 1% sẽ giúp bệnh nhân giảm 3% tiến triển thành ung thư đại tràng và nội soi đại tràng đã góp phần giảm 30% tỉ lệ mắc ung thư đại trực tràng [1]. Tuy nhiên, quá trình nội soi đại tràng yêu cầu thiết bị kỹ thuật đắt tiền và đội ngũ bác sĩ được đào tạo bài bản. Trong quá trình nội soi, một ống dài linh hoạt mềm và nhỏ gọn gọi là ống nội soi (colonoscope) đầu có gắn máy quay phim và đèn soi được đưa vào đại tràng, hình ảnh thu được trong quá trình nội soi được phóng đại trên màn hình màu có độ nét cao, cho phép bác sĩ xem xét bên trong của toàn bộ đại tràng. Chất lượng của quá trình nội soi phụ thuộc vào tay nghề, kinh nghiệm và sự tập trung của các bác sĩ nội soi. Các nghiên cứu gần đây chỉ ra rằng tỉ lệ polyp bị bỏ sót trong quá trình nội soi tương đối cao, dao động từ 20-47% bởi ba nguyên nhân chính: (i) Polyp có hình dạng, kích thước, màu sắc và kết cấu đa dạng ngay cả khi các polyp thuộc cùng một loại. (ii) Trong ảnh nội soi, đường biên ranh giới giữa polyp và vùng niêm mạc xung quanh nó không rõ ràng (hay vùng chứa polyp có màu sắc khá tương đồng với màu của vùng không chứa polyp). (iii) Trong quá trình nội soi, điều kiện ánh sáng khác nhau hay quá trình chuẩn bị ruột không sạch, xuất hiện bọt hoặc dịch nhầy cũng gây nên trở ngại lớn trong phát hiện polyp.

Do đó, ứng dụng công nghệ thông tin, đặc biệt là xây dựng các thuật toán trí tuệ nhân tạo trong phát hiện và phân vùng polyp trên ảnh nội soi là hướng đi cần thiết đặt ra cho ngành y học bởi những lợi ích to lớn trong việc hỗ trợ bác sĩ cải thiện độ chính xác, giảm thiểu việc bỏ sót polyp trong quá trình nội soi, đồng thời nâng cao năng lực của phòng khám trong khi vẫn duy trì được chất lượng chẩn đoán, sử dụng tối ưu các nguồn lực y tế còn đang thiếu hụt hiện nay.



Hình 1.2: Một ví dụ về: (a) ảnh nội soi đại tràng và (b) polyp được đánh nhãn bởi bác sĩ.

1.2 Các giải pháp hiện tại và hạn chế

Các phương pháp học máy truyền thống phân vùng polyp chủ yếu dựa trên các đặc trưng mức thấp như thông tin kết cấu, đặc điểm hình thái hoặc phân cụm điểm ảnh. Tuy nhiên, những phương pháp này thường có hiệu quả phân vùng thấp và khả năng tổng quát hóa không tốt. Với sự phát triển của việc ứng dụng học sâu trong phân tích ảnh y tế, bài toán phân vùng polyp ngày càng đạt được những kết quả hứa hẹn. Cụ thể, các kiến trúc có hình dạng chữ U kết hợp giữa mạng tích chập làm nhiệm vụ mã hóa và mạng giải chập làm nhiệm vụ giải mã đã đạt được hiệu năng đáng kể khi tận dụng được các đặc trưng ở nhiều mức khác nhau để tái tạo kết quả phân vùng có độ phân giải cao. PraNet [2] hướng tới xây dựng một kiến trúc mạng mới là sự kết hợp của 3 thành phần: (i) phần mã hóa nhằm trích xuất đặc trưng, (ii) phần giải mã nhằm đưa ra dự đoán ban đầu hình dạng của polyp và (iii) cơ chế chú ý ngược nhằm tinh chỉnh dự đoán. TransUNet [3], TransFuse [4], Polyp-PVT [5] là những nghiên cứu kết hợp những ưu điểm của CNN và Transformer nhằm nâng cao khả năng trích xuất và biểu diễn đặc trưng. Những kiến trúc này đạt được độ chính xác và khả năng tổng quát hóa vượt trội so với các phương pháp truyền thống. Tuy nhiên, bài toán phân vùng polyp vẫn còn tồn tại một số thách thức: (i) Polyp được hình thành từ các tế bào phát triển bất thường bên trong đại tràng, do đó màu sắc và kết cấu của vùng chứa polyp khá tương đồng với vùng niêm mạc xung quanh dẫn đến khó phân vùng chính xác, đặc biệt là vùng biên của polyp, (ii) Ảnh bị nhiễu trong quá trình nội soi hoặc quá trình thu thập dữ liệu do ống kính cần quay trong ruột để thu được ảnh polyp ở nhiều góc khác nhau dẫn đến có thể gây ra hiện tượng nhòe chuyển động và các vấn đề về độ phản xạ, làm tăng đáng kể độ khó trong việc phát hiện polyp, (iii) Có những trường hợp diện tích của polyp chỉ chiếm một phần rất nhỏ trong ảnh, (iv) Khả năng tổng quát hóa của mô hình trên các nguồn dữ liệu khác nhau.

1.3 Mục tiêu và định hướng giải pháp

Mục tiêu của đồ án là xây dựng một mô hình mạng nơron nhân tạo có khả năng phân vùng polyp tự động từ ảnh nội soi đại tràng với độ chính xác cao, có khả năng tổng quát hóa tốt trên các bộ dữ liệu đa dạng, đồng thời có kích thước lưu trữ nhỏ và độ phức tạp tính toán thấp.

Dựa trên những nghiên cứu gần đây: ViT [6], Segformer [7], DeiT [8], PVT [9], Swin [10] đã cho thấy kết quả đầy hứa hẹn khi áp dụng Transformer vào các tác vụ xử lý ảnh. Trong đồ án này sử dụng kiến trúc dựa trên Transformer là Mix Transformer để trích xuất đặc trưng ảnh, tạo ra đặc trưng ở nhiều mức và nhiều tỉ lệ khác nhau. Để kết hợp hiệu quả các đặc trưng này, đồ án đề xuất mô-đun giải mã

tích hợp song song các đặc trưng với sự trợ giúp của cơ chế chú ý theo cả chiều sâu và chiều không gian nhằm tạo bản đồ đặc trưng toàn cục có khả năng xác định được vị trí tương đối và hình dạng của polyp. Đồng thời sử dụng mô-đun chú ý ngược (Reverse Attention) và trích xuất đặc trưng kim tự tháp theo kênh (Channel-wise Feature Pyramid) nhằm cải thiện kết quả phân vùng tại biên của polyp và nâng cao độ chính xác khi phân vùng ảnh nội soi chứa polyp có kích thước nhỏ.

1.4 Đóng góp của đồ án

Đồ án này có 3 đóng góp chính như sau:

1. Tìm hiểu kiến trúc mô hình của các phương pháp đã có cho bài toán phân vùng polyp trên ảnh nội soi.
2. Đề xuất mô hình phân vùng polyp bao gồm kết hợp của ba thành phần: mô-đun mã hóa sử dụng kiến trúc Transformer làm bộ khung trích xuất đặc trưng thay vì CNN truyền thống; mô-đun giải mã giải mã kết hợp song song các đặc trưng với sự hỗ trợ của cơ chế chú ý theo cả chiều sâu và chiều không gian; mô-đun tinh chỉnh đặc trưng sử dụng cơ chế chú ý ngược áp dụng lên cả 4 đặc trưng cho ra tại 4 giai đoạn của kiến trúc Transformer.
3. Đánh giá và phân tích hiệu năng của mô hình đề xuất trên các tập dữ liệu mở về ảnh nội soi polyp đại trà.

1.5 Bố cục đồ án

Phần còn lại của báo cáo đồ án tốt nghiệp này được tổ chức như sau:

Chương 2 trình bày về cơ sở lý thuyết nền tảng, bao gồm tổng quan bài toán phân vùng ảnh, các kiến thức cơ bản của mạng nơron tích chập và các phương pháp tính tích chập phổ biến. Chương này cũng giới thiệu về Transformer truyền thống được ứng dụng cho bài toán xử lý ngôn ngữ tự nhiên và Vision Transformer được ứng dụng cho bài toán phân loại ảnh.

Chương 3 trình bày chi tiết kiến trúc mô hình đề xuất, bao gồm mô-đun mã hóa, mô-đun giải mã và mô-đun tinh chỉnh đặc trưng. Trong chương này cũng đưa ra hàm mất mát cần tối ưu, hàm mất mát này là sự kết hợp giữa Binary Cross-entropy và mean IoU có trọng số được đánh cho từng điểm ảnh.

Chương 4 trình bày về các độ đo đánh giá hiệu năng của mô hình, thông số các bộ dữ liệu sử dụng, phương pháp huấn luyện mô hình đề xuất, các phương pháp thực hiện thí nghiệm và các kết quả thực nghiệm định tính và định lượng.

Chương 5 tổng kết các kết quả đạt được trong đồ án, rút ra kết luận và định hướng hướng phát triển trong tương lai.

CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT

Trong chương 1 đã đưa ra các khó khăn, thách thức và hướng tiếp cận để giải quyết bài toán. Trong chương này sẽ trình bày một số khái niệm cơ bản, tổng quan về bài toán phân vùng ảnh, mạng nơron tích chập và một số phương pháp tính tích chập thông dụng được sử dụng trong các kiến trúc mạng giải quyết bài toán phân vùng ảnh. Trong chương 2 cũng giới thiệu kiến trúc Transformer truyền thống cho bài toán xử lý ngôn ngữ tự nhiên và việc áp dụng Transformer cho bài toán phân loại ảnh trong kiến trúc của Vision Transformer. Đồng thời, chương 2 cũng giới thiệu một vài nghiên cứu liên quan giải quyết bài toán phân vùng polyp trên ảnh nội soi.

2.1 Tổng quan bài toán phân vùng ảnh

Phân vùng ảnh (hay phân đoạn ảnh) là một phương pháp mà trong đó hình ảnh kỹ thuật số được chia thành nhiều nhóm con khác nhau gọi là segment. Hay nói một cách dễ hiểu, phân vùng ảnh là thực hiện phân chia một bức ảnh thành nhiều vùng khác nhau bằng việc gán nhãn cho từng điểm ảnh của bức ảnh, các điểm ảnh thuộc cùng một nhãn sẽ có những đặc tính giống nhau về màu sắc, cường độ hoặc kết cấu của ảnh. Đây là lớp bài toán quan trọng tạo bước tiền đề của quá trình xử lý dữ liệu hình ảnh. Mục tiêu của phân vùng ảnh là để đơn giản hóa hoặc thay đổi biểu diễn của bức ảnh, kết quả phân vùng tốt sẽ tạo điều kiện thuận lợi cho các khâu xử lý về sau, giúp cho quá trình phân tích bức ảnh trở nên đơn giản hơn. Thông qua phân vùng ảnh giúp ta hiểu được nội dung của một bức ảnh ở mức độ sâu khi biết được đồng thời: vị trí của vật thể trong ảnh, hình dạng của vật thể và từng điểm ảnh nào thuộc về vật thể nào.

Có hai dạng chính của bài toán phân vùng ảnh là phân vùng ngữ nghĩa (semantic segmentation) và phân vùng cá thể (instance segmentation). Trong đó, phân vùng ngữ nghĩa sẽ phân vùng ảnh theo những nhãn khác nhau mà không phân biệt sự khác nhau giữa các đối tượng trong cùng một nhãn; phân vùng cá thể sẽ phân vùng ảnh chi tiết đến từng đối tượng trong mỗi nhãn. Sự khác nhau giữa hai dạng bài toán này được minh họa ở Hình 2.1:



Hình 2.1: Minh họa sự khác biệt giữa bài toán phân vùng ngữ nghĩa và phân vùng cá thể¹.

Phân vùng ảnh có rất nhiều ứng dụng trong y học, xe tự hành, nông nghiệp hay trong xử lý ảnh vệ tinh, Chẳng hạn, trong:

Y học: phân vùng ảnh có thể hỗ trợ bác sĩ chẩn đoán khối u từ ảnh x-quang. Vì hình dạng các tế bào ung thư là một trong những yếu tố quyết định độ ác tính của bệnh, do đó phân vùng ảnh không những tìm ra vị trí, mà còn giúp biết được chính xác hình dạng của các tế bào ung thư để có phương pháp điều trị phù hợp và kịp thời. Ngoài ra, nó còn giúp tăng cường khả năng phân tích x-quang cho các chuyên gia, giúp việc chẩn đoán trở nên nhanh chóng và tiết kiệm thời gian.

Xe tự hành: đòi hỏi phải liên tục nhận thức, xử lý thông tin và lên kế hoạch trong một môi trường phát triển liên tục. Vì yêu cầu an toàn tuyệt đối và độ chính xác cao trong mọi quyết định nên một hệ thống xe tự hành cần xác định chính xác các vật thể xuất hiện khi tham gia giao thông như: người, đèn tín hiệu, xe cộ xung quanh, vạch kẻ đường hay biển báo giao thông, ... Phân vùng ảnh có thể cung cấp các thông tin về không gian trên đường di chuyển và phát hiện các đối tượng cần thiết theo yêu cầu của một hệ thống xe tự hành.

Ứng dụng trong nông nghiệp: hệ thống phun thuốc sâu tự động có khả năng phân biệt được diện tích cỏ và cây trồng dựa trên thuật toán phân vùng ảnh, do đó có thể tiết kiệm được lượng lớn thuốc trừ sâu, đồng thời khi diện tích cỏ lấn át so với diện tích cây trồng thì hệ thống có thể tự động được kích hoạt.

Xử lý ảnh vệ tinh: các vệ tinh nhân tạo quay quanh Trái Đất sẽ liên tục thu thập hình ảnh bề mặt Trái Đất ở những vùng khác nhau. Từ các bức ảnh vệ tinh này, mô hình phân vùng ảnh sẽ phân ảnh thành tuyến đường, khu phố, biển, cây cối, ...

Trên đây là một vài ứng dụng điển hình của thuật toán phân vùng ảnh. Còn rất nhiều các ứng dụng tiềm năng khác của thuật toán này vẫn đang được tiếp tục khai thác.

Bài toán phân vùng ảnh có thể được thực hiện thông qua 5 phương pháp chính:

- (i) Phân vùng dựa trên ngưỡng (Threshold Based Segmentation): là một dạng phân vùng ảnh đơn giản dựa trên việc đặt giá trị ngưỡng theo cường độ điểm ảnh của ảnh gốc. Để xác định giá trị ngưỡng, cần xem xét biểu đồ phân phối cường độ sáng của tất cả điểm ảnh trong ảnh, sau đó tiến hành chọn các ngưỡng để chia ảnh thành các phân vùng.

¹<https://developer.nvidia.com/blog/image-segmentation-using-digits-5/>

- (ii) Phân vùng dựa trên cạnh (Edge Based Segmentation): cạnh trong một hình ảnh là những vị trí không liên tục hay có sự biến đổi đột ngột về cường độ sáng, màu sắc hay kết cấu. Để phát hiện cạnh, ta có thể dùng các toán tử như Sobel, Laplace, Canny, ..., những toán tử này cung cấp thông tin cả về “độ lớn” và “hướng” của cạnh. Cuối cùng, để thu được kết quả phân vùng cần thực hiện một số bước bổ sung bao gồm: liên kết các cạnh liền kề và kết hợp chúng để hình thành nên phân vùng chứa đối tượng.
- (iii) Phân vùng dựa trên khu vực (Region-Based Segmentation): một vùng hay khu vực được định nghĩa là một nhóm các điểm ảnh kết nối với nhau và có các thuộc tính tương đồng về cường độ, màu sắc,... Phương pháp này làm việc khá hiệu quả trong trường hợp ảnh bị nhiễu, có thể được thực hiện thông qua hai phương pháp chính: phát triển khu vực (region growing method) hoặc phân tách và hợp nhất khu vực (region splitting and merging method).
- (iv) Phân vùng dựa trên kỹ thuật phân cụm (Clustering Based Segmentation): phân cụm là một thuật toán học máy không giám sát được sử dụng phổ biến trong phân vùng ảnh. Một trong những thuật toán phân cụm thường được ứng dụng cho bài toán phân vùng ảnh là K-means. Thuật toán này sẽ phân cụm cường độ điểm ảnh trên ảnh thành K cụm, sau đó giá trị của mỗi điểm ảnh này sẽ được thay thế bởi giá trị tâm cụm nhằm phân vùng hình ảnh. Điểm hạn chế của thuật toán này là tốn kém chi phí tính toán, vì khi huấn luyện cần tính khoảng cách từ tâm cụm tới toàn bộ các điểm ảnh trong ảnh, đồng thời siêu tham số K được chọn là không chắc chắn.
- (v) Phân vùng dựa trên mạng nơron nhân tạo (Artificial Neural Network Based Segmentation): phương pháp này sử dụng trí tuệ nhân tạo để tự động phân tích hình ảnh. Mạng nơron tích chập được sử dụng khá phổ biến trong phân vùng ảnh vì chúng có thể xác định, xử lý dữ liệu hình ảnh một cách nhanh chóng, hiệu quả và mang lại độ chính xác cao nhờ tận dụng được ưu thế của dữ liệu lớn và khả năng trích chọn đặc trưng tự động.

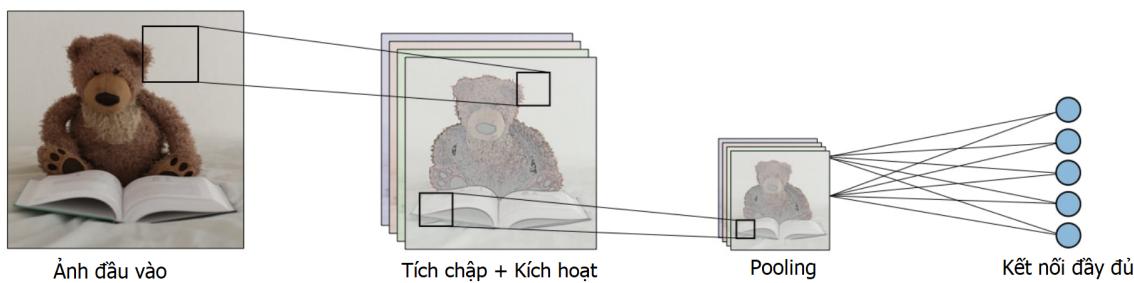
2.2 Mạng nơron tích chập

2.2.1 Tổng quan mạng nơron tích chập

Mạng nơron tích chập (Convolutional neural networks), còn được biết đến với tên CNN, được sử dụng lần đầu tiên vào cuối những năm 1998 bởi Yann LeCun và các cộng sự cho bài toán phân loại chữ số viết tay và đạt được hiệu quả cao. Với sự bùng nổ của dữ liệu và khả năng tính toán của phần cứng, CNN đã dần được sử dụng để thay thế các phương pháp tạo đặc trưng thủ công và các thuật toán học máy truyền thống, ngày càng được ứng dụng phổ biến trong nhiều lĩnh vực như thị

giác máy tính, hệ thống gợi ý, xử lý ngôn ngữ tự nhiên, xử lý âm thanh và phân tích dữ liệu chuỗi thời gian.

Về cơ bản thiết kế của một mạng nơron tích chập bao gồm 5 thành phần như được minh họa trong Hình 2.2: (i) ảnh đầu vào, (ii) tầng tích chập sử dụng các bộ lọc để trích xuất đặc trưng, (iii) tầng kích hoạt giúp tăng tính phi tuyến cho mạng, (iv) tầng gộp (pooling) giúp giảm chi phí tính toán, đồng thời tăng tính bất biến của không gian và (v) tầng kết nối đầy đủ nhằm tìm ra phân phối xác suất, tối ưu hóa mục tiêu của mạng.



Hình 2.2: Kiến trúc cơ bản của một mạng nơron tích chập².

Trong mạng nơron tích chập, trọng số của các bộ lọc được học một cách tự động. Mục tiêu của phép tích chập là trích xuất đặc trưng như các đường, cạnh, thông tin kết cấu,... từ ảnh đầu vào. Mạng nơron tích chập không nhất thiết chỉ giới hạn trong một lớp tích chập. Thông thường, lớp tích chập đầu tiên chịu trách nhiệm nắm bắt các đặc trưng mức thấp như màu sắc, hướng dốc gradient,... với các lớp tích chập được thêm vào, mô hình sẽ nắm bắt các đặc trưng ngữ nghĩa cấp cao, mang đến cho chúng ta một mạng lưới nơron tích chập có sự hiểu biết toàn diện về hình ảnh trong bộ dữ liệu, tương tự như cách con người nhận về hình ảnh.

Mạng nơron tích chập có hai tính chất quan trọng là: kết nối cục bộ và chia sẻ tham số. Khác với các mạng nơron thông thường, mạng nơron tích chập không kết nối tới toàn bộ hình ảnh mà chỉ kết nối tới từng vùng địa phương (local region) hoặc vùng nhận thức (receptive field) có kích thước bằng kích thước bộ lọc tích chập. Do đó, mỗi thành phần đầu ra chỉ phụ thuộc vào bộ phát hiện đặc trưng và một phần nhỏ của ảnh đầu vào thay vì toàn bộ bức ảnh. Các vùng địa phương này sẽ được chia sẻ chung một bộ tham số có tác dụng nhận thức đặc trưng của bộ lọc. Nhờ những tính chất này mà mạng nơron tích chập có nhiều ưu điểm nổi bật so với mạng nơron thông thường. Bên cạnh hiệu năng cao trên số lượng mẫu cần thiết để đạt được độ chính xác, các mạng nơron tích chập thường có hiệu quả tính toán hơn, bởi đòi hỏi ít tham số và dễ thực thi song song trên nhiều GPU hơn các kiến trúc mạng kết nối đầy đủ.

²<https://stanford.edu/~shervine/lvi/teaching/cs-230/cheatsheet-convolutional-neural-networks>

2.2.2 Một số phương pháp tính tích phân

Tích chập là một khái niệm đóng vai trò quan trọng và xuất hiện sớm trong xử lý tín hiệu số nhằm biến đổi thông tin đầu vào thông qua một phép tích chập với bộ lọc để trả về đầu ra là một tín hiệu mới. Tín hiệu mới này sẽ làm giảm những đặc trưng mà bộ lọc không quan tâm và chỉ giữ lại những đặc trưng chính quan trọng. Trong phần này sẽ trình bày một số phương pháp tính tích chập phổ biến được áp dụng cho bài toán phân vùng ảnh.

a, Tích chập thông thường

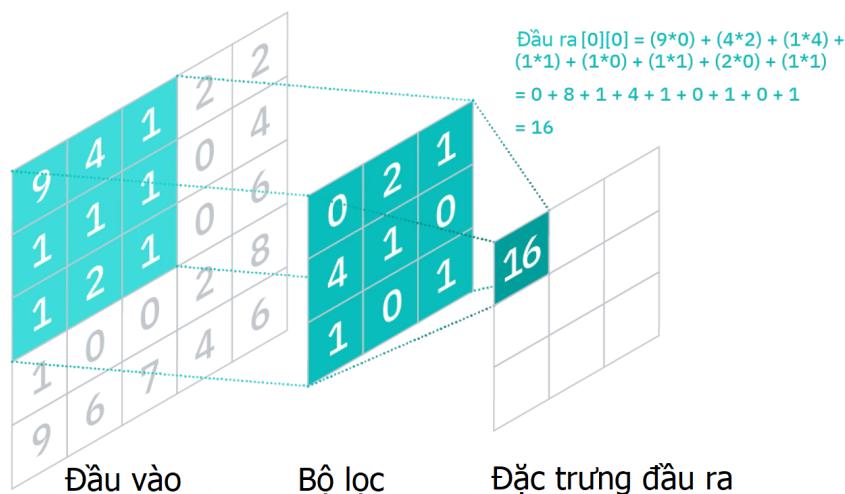
Tích chập thông dụng nhất là tích chập hai chiều, được áp dụng trên ma trận đầu vào và ma trận bộ lọc hai chiều (xem hình minh họa 2.3). Kết quả của phép tích chập giữa một ma trận $X \in R^{H \times W}$ với một bộ lọc $F \in R^{F \times F}$ là một ma trận $Y \in R^{A \times B}$ sẽ trải qua những bước sau:

Bước 1: tính tích chập tại một điểm: tại vị trí đầu tiên trên cùng của ma trận đầu vào, ta sẽ trích ra một ma trận con $X_{sub} \in R^{F \times F}$ có kích thước bằng với kích thước của bộ lọc. Giá trị y_{11} trên Y là tích chập của X_{sub} với F, được tính như sau: $y_{11} = \sum_{i=1}^F \sum_{j=1}^F x_{ij} f_{ij}$ (trong đó: chỉ số ma trận được quy ước bắt đầu từ 1, $x_{ij} \in X_{sub}$ và $f_{ij} \in F$).

Bước 2: tiến hành trượt bộ lọc theo chiều từ trái qua phải, từ trên xuống dưới với bước nhảy S, ta sẽ tính được các giá trị y_{ij} tiếp theo.

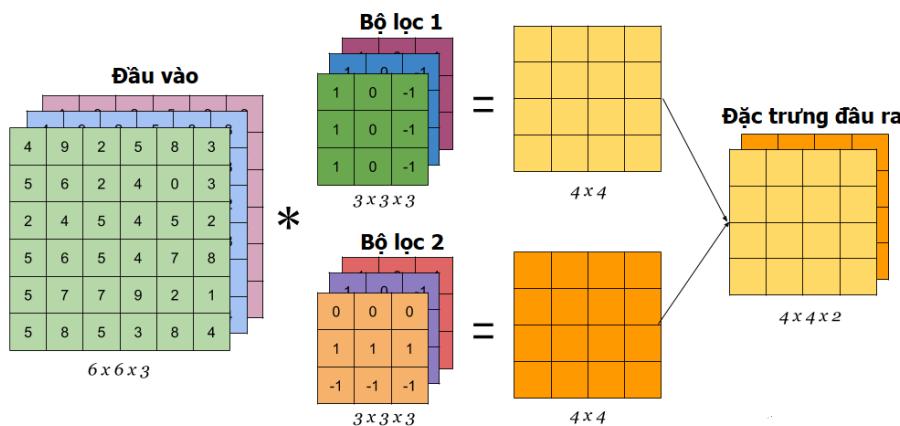
Nếu kích thước của ma trận đầu vào là $n_h \times n_w$ và kích thước của bộ lọc tích chập là $f_h \times f_w$, kích thước của đặc trưng đầu ra sẽ là:

$$(n_h - f_h + 1) \times (n_w - f_w + 1)$$



Hình 2.3: Minh họa tính tích chập hai chiều³.

Trong trường hợp ảnh đầu vào có 3 kênh RGB, phép tính tích chập cũng được thực hiện một cách tương tự nhưng thay vì 2 chiều, bộ lọc được sử dụng sẽ có 3 chiều. Hình 2.4 minh họa quá trình thực hiện tích chập 3 chiều. Trong đó, ảnh đầu vào là một ma trận 3 chiều có kích thước $6 \times 6 \times 3$, ta áp dụng lên ảnh hai bộ lọc, mỗi bộ lọc có kích thước $3 \times 3 \times 3$ và bước nhảy tính chập bằng 1 để thu được hai đặc trưng có kích thước 4×4 . Hai đặc trưng này được kết hợp với nhau để thu được đặc trưng đầu ra cuối cùng có kích thước $4 \times 4 \times 2$.



Hình 2.4: Minh họa tính tích chập 3 chiều ⁴.

b, Tích chập dãn nở (Dilation Convolution)

Tích chập dãn nở là một kỹ thuật thực hiện bỏ qua các điểm ảnh của ma trận đầu vào một cách xen kẽ, hay nói cách khác, phương pháp này thực hiện làm dãn nở bộ lọc bằng cách chèn thêm khoảng trống giữa các phần tử của bộ lọc trong khi thực hiện phép tích chập. Khi tỉ lệ dãn nở bằng 1, tích chập dãn nở chính là tích chập thông thường. Hình 2.5 minh họa tích chập dãn nở với tỉ lệ dãn nở bằng 2.

Phép tích chập này giúp mở rộng trường tiếp nhận cục bộ, lấy được nhiều thông tin hơn từ ảnh đầu vào mà không tăng độ phức tạp tính toán. Tích chập dãn nở có thể được sử dụng để thay thế phép gộp (pooling), giúp giảm bộ nhớ tiêu thụ và không làm mất độ phân giải của đặc trưng đầu ra.

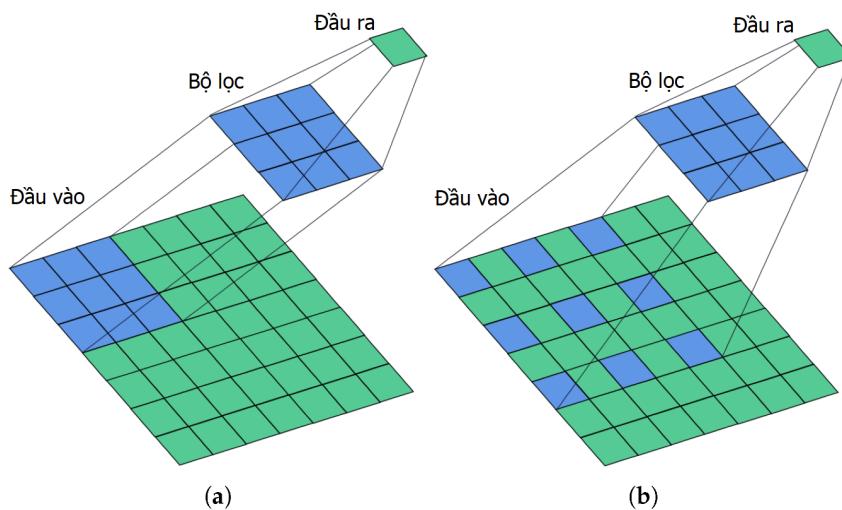
c, Tích chập tách biệt chiều sâu (Depth-wise convolution)

Độ sâu là một trong những nguyên nhân chính dẫn tới sự gia tăng số lượng tham số của mô hình. Tích chập tách biệt chiều sâu sẽ tìm cách loại bỏ sự phụ thuộc vào độ sâu khi tích chập mà vẫn tạo ra được đặc trưng đầu ra có kích thước tương đương so với tích chập thông thường. Để làm được điều này, tích chập tách biệt chiều sâu

³<https://www.ibm.com/cloud/learn/convolutional-neural-networks>

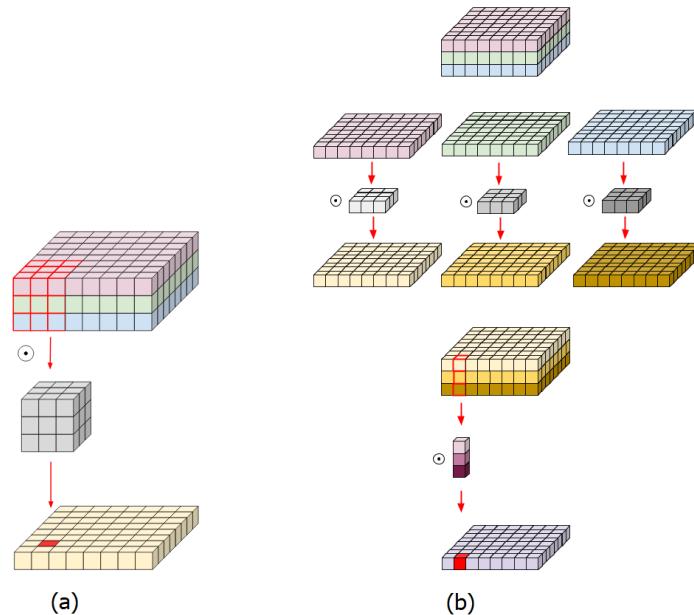
⁴<https://indoml.com/2018/03/07/student-notes-convolutional-neural-networks-cnn-introduction/>

⁵<https://www.mdpi.com/1424-8220/21/21/7319/htm>



Hình 2.5: (a) Phép tích chập thông thường, (b) Phép tích chập dãn nở với tỉ lệ dãn nở bằng hai⁵.

sẽ áp dụng lên mỗi kênh của ma trận đầu vào một bộ lọc tích chập khác nhau và hoàn toàn không chia sẻ trọng số. Kết quả đầu ra sẽ áp dụng tích chập điểm với bộ lọc kích thước 1×1 để thu được ma trận đặc trưng có chiều sâu mong muốn. Hình 2.6 minh họa sự khác nhau giữa tích chập thông thường và tích chập tách biệt chiều sâu.



Hình 2.6: (a) Phép tích chập thông thường, (b) Phép tích chập tách biệt chiều sâu⁶.

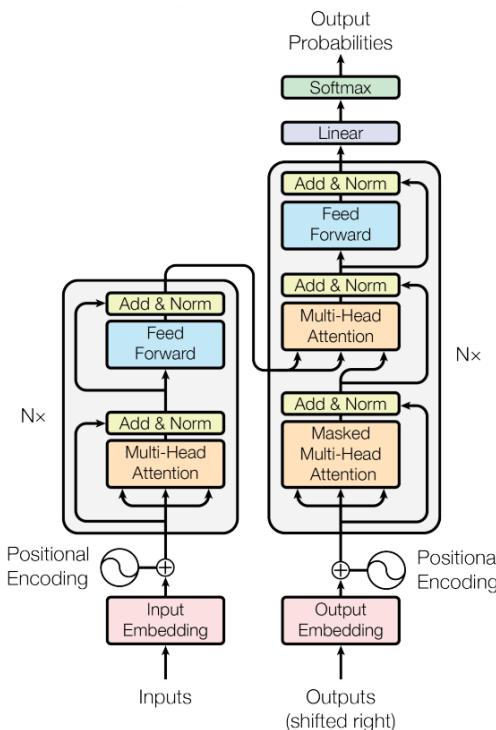
Tích chập tách biệt chiều sâu có một số ưu điểm như: (i) khả năng nhận diện đặc trưng tốt hơn bởi quá trình học được tách biệt theo từng kênh và từng bộ lọc.

⁶<https://medium.com/@zurister/depth-wise-convolution-and-depth-wise-separable-convolution-37346565d4ec>

Nếu đặc trưng trên các kênh là khác nhau thì sử dụng các bộ lọc riêng cho kênh sẽ chuyên biệt hơn trong việc phát hiện các đặc trưng. (ii) Giảm thiểu số lượng tham số: giả sử số kênh của đầu vào là c , kích thước bộ lọc là $k \times k$, số kênh đầu ra mong muốn là c' . Với tích chập thông thường, số tham số cần sử dụng là $k \times k \times c \times c'$. Trong khi đó, với tích chập tách biệt chiều sâu, số tham số cần sử dụng là $k \times k \times c + c \times c'$, ít hơn gấp gần c' lần so với tích chập thông thường.

2.3 Transformer truyền thống

Transformer ban đầu được thiết kế cho bài toán xử lý ngôn ngữ tự nhiên, được các kỹ sư Google giới thiệu vào năm 2017 trong bài báo Attention Is All You Need [11] nhằm giải quyết những hạn chế của các kiến trúc mạng hồi quy trước đó như RNN hay LSTM là khó bắt được sự phụ thuộc xa giữa các từ trong câu và tốc độ huấn luyện chậm do phải xử lý đầu vào tuần tự. Kiến trúc Transformer ra đời đã tạo nên bước đột phá mới trong công nghệ xử lý ngôn ngữ tự nhiên.



Hình 2.7: Kiến trúc Transformer truyền thống [11].

Kiến trúc Transformer (xem Hình 2.7) gồm phần mã hóa (nhánh bên trái) và giải mã (nhánh bên phải):

- Nhánh mã hóa: nhánh này nhằm mục đích chuyển đầu vào thô thành những đặc trưng có thể được học bởi mô hình. Nhánh mã hóa được xây dựng bao gồm xếp chồng lên nhau của 6 khối. Mỗi khối lại bao gồm hai tầng con: chú ý đa đầu (multi-head self-attention) và truyền xuôi kết nối đầy đủ (fully-connected feed-forward).

- Nhánh giải mã: nhận đầu vào là kết quả của nhánh mã hóa, nhánh này nhằm mục đích tìm ra phân phối xác suất từ các đặc trưng học được từ nhánh mã hóa để xác định kết quả đầu ra. Nhánh giải mã cũng là tổng hợp xếp chồng của 6 khối. Kiến trúc tương tự như các khối của nhánh mã hóa, ngoại trừ tầng con đầu tiên được thiết kế để mô hình không nhìn thấy được các từ trong tương lai.
- Vị trí và thứ tự của các từ là những phần quan trọng thiết yếu của bất kỳ ngôn ngữ nào. Tuy nhiên, Transformer xử lý các từ song song, toàn bộ câu đầu vào được truyền vào mô hình cùng một thời điểm, do đó bản thân mô hình không có bất kỳ thông tin nào về vị trí thứ tự cho từng từ. Vì vậy, cần có cơ chế mã hóa để đưa thông tin thứ tự của các từ vào mô hình Transformer và trong bài báo, nhóm tác giả gọi cách mã hóa đó là mã hóa vị trí (position encoding). Để tính giá trị mã hóa vị trí tại chiều vector tương ứng, tại vị trí chẵn tác giả sử dụng hàm sin, tại vị trí lẻ tác giả sử dụng hàm cos theo công thức:

$$p_t^i = f(t)^i = \begin{cases} \sin(w_k \times t), \text{ nếu } i = 2k \\ \cos(w_k \times t), \text{ nếu } i = 2k + 1 \end{cases}$$

Trong đó:

- t : chỉ số của từ hiện tại
- i : chỉ số của chiều trong vector mã hóa
- d : kích thước của vector mã hóa
- $w_k = \frac{1}{10000^{2k/d}}$

Cơ chế chú ý (self-attention): cơ chế này chính là một phần cốt yếu đóng góp nêu sự thành công của Transformer thông qua việc giúp mô hình tăng khả năng tập trung hay chú ý vào những phần quan trọng, trong khi làm mờ đi những thông tin không liên quan. Với bài toán dịch máy sử dụng mạng nơron hồi tiếp, mô-đun mã hóa phải nén tất cả thông tin của một câu thành một vector biểu diễn duy nhất, chứa toàn bộ thông tin cần thiết để mô-đun giải mã có thể dịch thành câu đích. Tuy nhiên, những câu dài sẽ không được dịch chính xác vì thông tin không được lưu trữ đủ trong một vector biểu diễn duy nhất. Cũng tương tự như trong bài toán sinh mô tả cho ảnh, mô hình CNN cần lưu trữ thông tin của toàn bộ bức ảnh thành một vector duy nhất, rồi sau đó mô-đun giải mã sẽ phát sinh câu mô tả dựa theo thông tin lưu trong vector này, tuy nhiên, vector này có thể không chứa đủ thông tin để phát sinh cho toàn bộ câu mô tả. Để giải quyết vấn đề này, sử dụng cơ chế chú ý cho phép mô hình có thể tập trung chú ý vào từng phần của câu hoặc bức ảnh một cách rõ ràng, từ đó thông tin không cần phải nén vào một vector biểu diễn duy nhất.

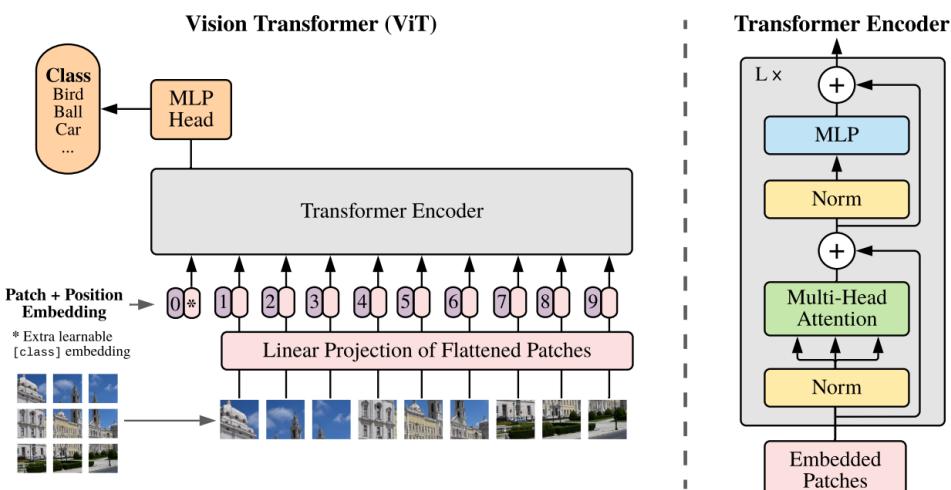
Quá trình tính ma trận chú ý được tổng hợp trong công thức:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Trong đó: đầu vào bao gồm ba ma trận Q (query), K (key) và V (value). Phép nhân hai ma trận $Q \times K^T$ sẽ cho ra ma trận trọng số chú ý (mức độ chú ý giữa các cặp từ). Sau đó, chia cho $\sqrt{d_k}$ nhằm mục đích tránh tràn số và sử dụng hàm softmax để tính phân phối xác suất cho ma trận trọng số chú ý. Cuối cùng, nhân kết quả với ma trận V để đưa ra vector đại diện cuối cùng.

2.4 Vision Transformer (ViT)

Mạng nơron tích chập đã đạt được những thành công đáng kể trong lĩnh vực thị giác máy tính, tuy nhiên các nghiên cứu gần đây chỉ ra rằng kiến trúc mạng Transformer và cơ chế chú ý cũng hoạt động tốt cho xử lý ảnh, hơn thế nữa còn mang lại hiệu năng vượt trội trong nhiều tác vụ. Trong đó ViT [6] là nghiên cứu tiên phong chứng minh rằng kiến trúc Transformer hoàn toàn có thể thay thế cho CNN truyền thống trong tác vụ phân loại ảnh.



Hình 2.8: Kiến trúc Vision Transformer [6]

ViT được thiết kế cho bài toán phân loại ảnh, kiến trúc mạng bao gồm 3 phần chính như được minh họa trong Hình 2.8: (i) Phép chiếu tuyến tính (Linear Projection of Flattened Patches). (ii) Mô-đun mã hóa là kiến trúc thuần Transformer. (iii) Mô-đun giải mã là khối perceptron đa tầng.

Với các mô hình CNN truyền thống cho bài toán phân loại ảnh, đầu vào cho mô hình là toàn bộ ảnh với kích thước cố định. Tuy nhiên ViT có cách xử lý khác, với mỗi ảnh đầu vào ViT chia ảnh ra thành các phần nhỏ có kích thước bằng nhau gọi là các patch. ViT xử lý các phần này như các từ trong một câu bằng cách đưa các

phần qua một tầng tuyến tính để thu được vector nhúng cho từng phần. Tiếp theo đó, đưa qua mô-đun mã hóa của Transformer để thu được vector bối cảnh (context vector) và cuối cùng, đưa qua khối perceptron đa tầng để thu được xác suất đầu ra tương ứng với các lớp.

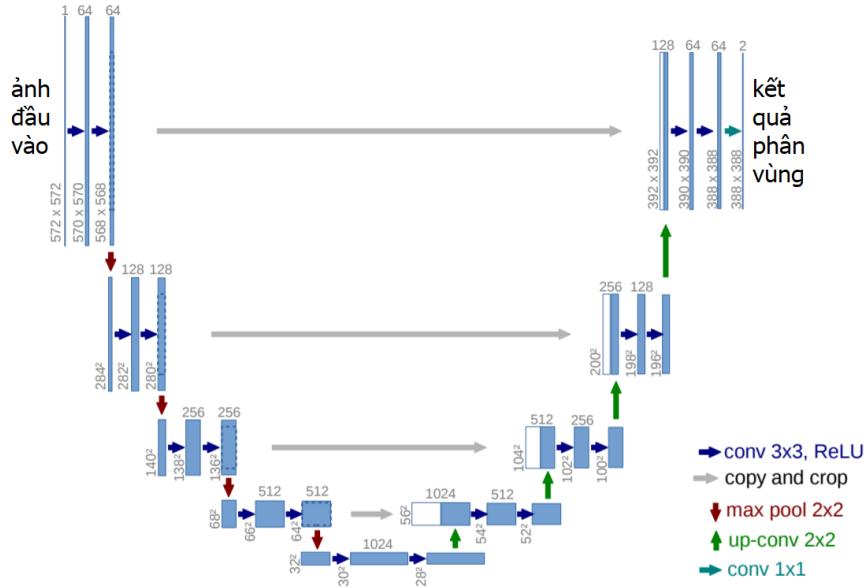
Kết quả thực nghiệm đã chứng minh phương pháp xử lý trong ViT giúp giảm đáng kể nguồn lực tính toán và đạt độ chính xác vượt trội so với mô hình CNN hiện đại tương tự (chẳng hạn như ResNet). Tuy nhiên, nhược điểm chính của ViT là yêu cầu bộ dữ liệu có kích thước tương đối lớn để duy trì hiệu năng, trong khi đó, các bộ dữ liệu mở về y tế thường có kích thước khá nhỏ, chẳng hạn bộ dữ liệu Kvasir chỉ chứa 1000 ảnh về nội soi đại tràng dù đây là bộ dữ liệu mở có kích thước lớn nhất cho bài toán phân vùng polyp.

2.5 Một số nghiên cứu liên quan đến bài toán phân vùng polyp

Phương pháp phân vùng polyp truyền thống: các phương pháp truyền thống phân vùng polyp chủ yếu dựa trên các đặc trưng mức thấp như thông tin kết cấu [12], đặc điểm hình thái [13] hoặc phân cụm điểm ảnh [14]. Tuy nhiên, do đặc điểm về độ đa dạng của dữ liệu và độ tương đồng cao giữa vùng có chứa polyp và vùng niêm mạc xung quanh nó, những phương pháp truyền thống này có hiệu quả thấp và tính tổng quát hóa không cao, mô hình thường phát hiện thiếu hoặc sai với xác suất lớn.

Phương pháp phân vùng polyp dựa trên học sâu: Với sự phát triển của việc ứng dụng học sâu trong phân tích ảnh y tế, bài toán phân vùng polyp ngày càng đạt được những kết quả hứa hẹn. Hầu hết các phương pháp phổ biến phân vùng polyp đại tràng trong những năm gần đây đều sử dụng mạng nơron tích chập. U-Net [15] là một kiến trúc CNN tiên phong trong xử lý ảnh y tế, kiến trúc này là sự kết hợp của hai nhánh: nhánh mã hóa và nhánh giải mã. Nhánh mã hóa có nhiệm vụ trích xuất đặc trưng đa mức từ ảnh đầu vào. Kết quả của nhánh mã hóa được đưa vào nhánh giải mã để tiếp tục trích xuất đặc trưng ở mức cao hơn, đồng thời tăng dần độ phân giải về kích thước ảnh gốc ban đầu để sinh ra nhãn phân vùng. Đầu ra các lớp của nhánh mã hóa là các đặc trưng mức thấp được chuyển trực tiếp sang lớp tương ứng của nhánh giải mã qua các kết nối tắt. Nhờ vậy, thông tin giàu ngữ cảnh ở nhánh giải mã sẽ được kết hợp với các thông tin chi tiết ở mức thấp. Hai nhánh này được kết nối theo hình dạng chữ U nên kiến trúc mạng có tên U-Net (xem Hình 2.9). Kế thừa những ưu điểm của U-Net, một loạt các biến thể như: UNet++ [16], DoubleUnet [17], ResUNet++ [18] được cải tiến để mang lại hiệu năng cao hơn. Những kiến trúc này đã dần thay thế hoàn toàn các phương pháp phân vùng truyền thống bởi mang lại hiệu năng vượt trội. Tuy nhiên, kết quả thực nghiệm cho

thấy kiến trúc mã hóa – giải mã U-Net cho kết quả không tốt khi phân vùng các đối tượng có kích thước nhỏ, đặc biệt trong ảnh nội soi đại tràng, diện tích polyp thường chiếm một tỉ lệ rất nhỏ trong ảnh.



Hình 2.9: Kiến trúc mạng U-Net [15].

Khác với các phương pháp dựa trên U-Net, kiến trúc PraNet [2] sử dụng mô-đun giải mã từng phần để tích hợp các đặc trưng ở mức ngữ nghĩa cao nhằm tạo ra bản đồ đặc trưng toàn cục (global feature map), đồng thời sử dụng mô-đun chú ý ngược nhằm tinh chỉnh đặc trưng và khai thác thêm thông tin về biên của polyp, bổ sung những thông tin này vào bản đồ đặc trưng toàn cục để tạo ra kết quả dự đoán cuối cùng. HarDNet-MSEG [19], CaraNet [20] là hai kiến trúc khác được nghiên cứu để cải tiến độ chính xác của PraNet. Trong đó, HarDNet-MSEG đề xuất kiến trúc mã hóa và giải mã đơn giản hơn, thay thế bộ khung trích xuất đặc trưng (backbone) Res2Net bởi HardNet mang lại độ chính xác cao hơn, đồng thời xóa bỏ cơ chế chú ý trong kiến trúc nhằm cải thiện tốc độ suy luận. Tuy có tốc độ suy luận thời gian thực ngay cả trên các thiết bị biên phổ thông nhưng độ chính xác của PraNet và HardNet-MSEG vẫn còn hạn chế trên các ảnh polyp có kích thước nhỏ. Nhóm tác giả CaraNet đã cải tiến PraNet bằng cách tích hợp thêm mô-đun trích xuất đặc trưng kim tự tháp theo kênh (CFP) nhằm biểu diễn đặc trưng ở nhiều tỉ lệ khác nhau và cơ chế chú ý theo trực (axial attention) trong mô-đun chú ý ngược để nâng cao độ chính xác hơn nữa trong phát hiện biên của polyp. Nhờ các đóng góp này, CaraNet đạt độ chính xác vượt xa các phương pháp tiên tiến nhất hiện tại trên đồng thời bộ dữ liệu ảnh polyp thông thường và bộ dữ liệu ảnh polyp có kích thước nhỏ.

TransFuse [4] là kiến trúc mạng mới được thiết kế để tận dụng và kết hợp những ưu điểm của CNN và Transformer. TransFuse trích xuất đặc trưng đồng thời từ

hai nhánh mã hóa song song là CNN (ResNet) và Transformer (DeiT), nhờ đó TransFuse có thể trích xuất được cả đặc trưng ở mức cục bộ và đặc trưng ở mức toàn cục. Các đặc trưng này sẽ được kết hợp với nhau một cách hiệu quả thông qua mô-đun BiFusion tạo nên đặc trưng mới có khả năng biểu diễn mạnh mẽ. Gần đây, đồng nhóm tác giả của PraNet cũng xây dựng một kiến trúc mạng mới có tên Polyp-PVT [5] cho bài toán phân vùng polyp. Kiến trúc này sử dụng bộ khung PVTv2 để trích xuất đặc trưng, bộ khung này là một kiến trúc Transformer, được thiết kế theo cấu trúc kim tự tháp nhằm trích xuất đặc trưng ở nhiều mức và nhiều tỉ lệ khác nhau (multi-scale feature). Đồng thời, trong bài báo cũng thiết kế mô-đun giải mã để kết hợp hiệu quả các đặc trưng cho ra bởi phần mã hóa. Mô-đun giải mã này bao gồm 3 phần: (i) CFM - thu thập thông tin ngữ nghĩa và tìm ra vị trí tương đối của polyp thông qua việc tích hợp liên tục các đặc trưng mức cao. (ii) CIM - tìm ra thông tin quan trọng từ đặc trưng mức thấp như thông tin kết cấu, màu sắc hay cạnh nhờ sử dụng cơ chế chú ý thông qua mô-đun CBAM; (iii) SAM - mô-đun này kết hợp các đặc trưng mức thấp cho ra bởi CIM và đặc trưng mức cao cho ra bởi CFM sử dụng graph convolutional để tạo ra kết quả dự đoán cuối cùng. Những phương pháp này đều cho kết quả thực nghiệm vượt trội đáng kể so với các phương pháp trước đó, điều này cho thấy sự hiệu quả và những hướng đi đầy triển vọng khi áp dụng kiến trúc Transformer vào các tác vụ thị giác máy tính.

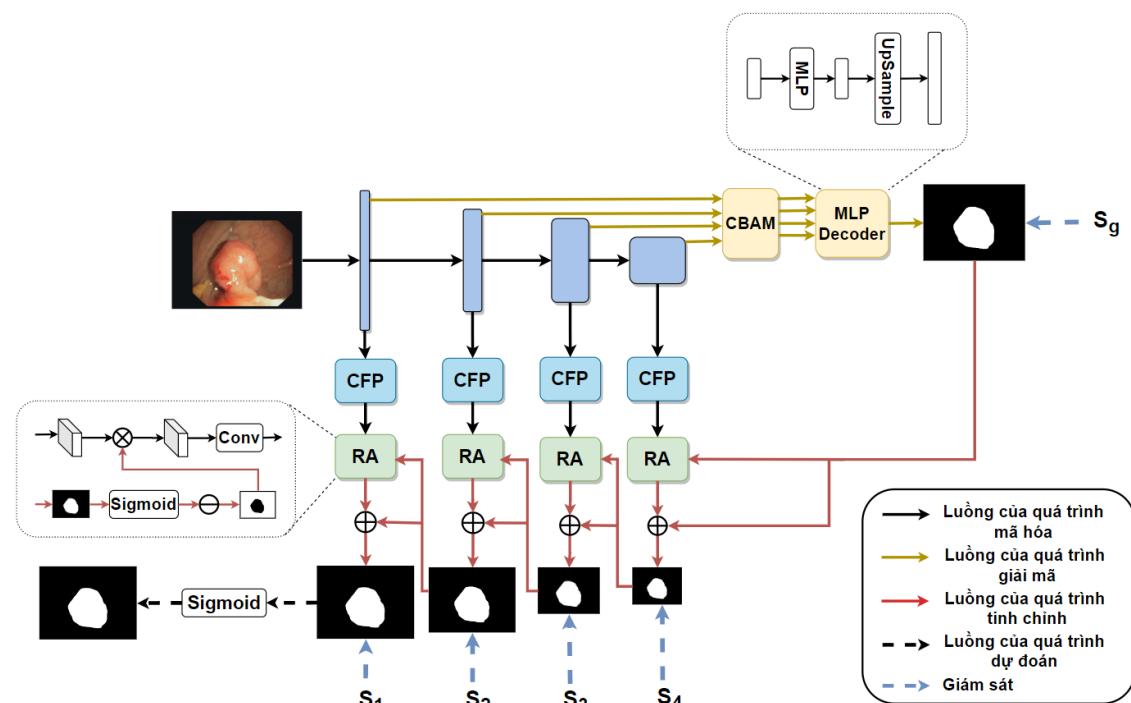
Kết chương: trong chương này đã trình bày khái niệm và các phương pháp cơ bản để giải quyết bài toán phân vùng ảnh, giới thiệu 3 phương pháp tính tích chập bao gồm: tích chập thông thường, tích chập dãn nở và tích chập tách biệt chiều sâu. Các phương pháp tích chập này được sử dụng phổ biến trong các kiến trúc mạng nơron cho bài toán phân vùng ảnh. Từ kết quả phân tích các nghiên cứu liên quan cho thấy tiềm năng của việc ứng dụng Transformer vào các tác vụ thị giác máy tính, trong đồ án này cũng sử dụng bộ khung dựa trên Transformer là Mix Transformer nhằm trích xuất đặc trưng ảnh. Đồng thời, tích hợp vào mô hình các mô-đun như CBAM và chú ý ngược nhằm tinh chỉnh đặc trưng, tăng độ chính xác dự đoán. Phần này được trình bày chi tiết trong chương 3.

CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

Trong chương 2 đã trình bày các kiến thức cơ sở về bài toán phân vùng ảnh. Chương này sẽ trình bày chi tiết 3 mô-đun được tích hợp trong mô hình đề xuất, bao gồm: mô-đun mã hóa, mô-đun giải mã và mô-đun tinh chỉnh đặc trưng. Cuối cùng, sẽ định nghĩa hàm mất mát mà mô hình cần tối ưu.

3.1 Tổng quan kiến trúc mô hình đề xuất

Như được thể hiện trong Hình 3.1, kiến trúc mô hình đề xuất bao gồm kết hợp của 3 thành phần: mô-đun mã hóa, mô-đun giải mã và mô-đun tinh chỉnh đặc trưng:



Hình 3.1: Tổng quan kiến trúc đề xuất.

Cụ thể, mô-đun mã hóa sử dụng kiến trúc Mix Transformer (MiT) nhằm trích xuất đặc trưng toàn cục và mô hình hóa sự phụ thuộc xa của đặc trưng ở nhiều tỉ lệ khác nhau từ ảnh đầu vào. Mô-đun giải mã tích hợp các đặc trưng cho ra bởi MiT nhằm tạo ra bản đồ đặc trưng toàn cục xấp xỉ hình dạng và vị trí của polyp. Mô-đun tinh chỉnh đặc trưng được thiết kế để tinh chỉnh bản đồ đặc trưng toàn cục, khám phá các thông tin còn thiếu, đặc biệt là vùng biên của polyp.

3.2 Mô-đun mã hóa

Dựa trên những nghiên cứu gần đây: ViT [6], Segformer [7], DeiT [8], PVT [9], Swin [10] đã cho thấy kết quả đầy hứa hẹn khi áp dụng Transformer vào các tác vụ xử lý ảnh. Trong đồ án này, thay vì sử dụng các kiến trúc CNN truyền thống, sẽ sử dụng kiến trúc dựa trên Transformer là Mix Transformer (MiT) để trích xuất

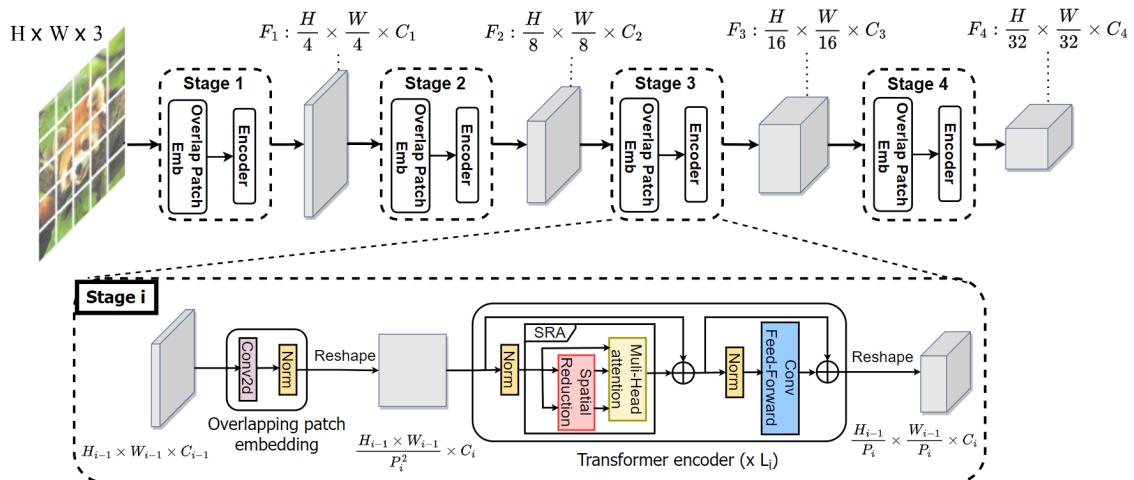
đặc trưng ảnh, tạo ra đặc trưng ở nhiều mức và nhiều tỉ lệ khác nhau. Kiến trúc này được đề xuất trong bài báo SegFormer [7].

Như đã trình bày trong Mục 2.4, kiến trúc ViT chỉ áp dụng cho bài toán phân loại ảnh, khó áp dụng trực tiếp cho các bài toán phân loại ở mức độ điểm ảnh như tác vụ phát hiện vật thể (object detection) hoặc phân vùng ảnh (segmentation). Bên cạnh đó, ViT cũng tồn tại một số nhược điểm như: các đặc trưng cho ra có kích thước chiều không gian giống nhau (single-scale) và có độ phân giải thấp bởi ViT chia ảnh đầu vào thành các phần có kích thước lớn (16x16). Đồng thời, chi phí tính toán và bộ nhớ lưu trữ tương đối lớn với ảnh đầu vào có kích thước phổ biến.

Để khắc phục những nhược điểm này, MiT là một kiến trúc Transformer được nghiên cứu để làm bộ khung trích xuất đặc trưng thay thế CNN truyền thống ở nhiều tác vụ phía sau, bao gồm dự đoán ở cả mức ảnh cũng như dự đoán ở mức điểm ảnh, bằng các đóng góp: (i) Chia ảnh thành các phần có kích thước nhỏ hơn (chẳng hạn 4x4 điểm ảnh ở mỗi phần) nhằm học được khả năng biểu diễn có độ phân giải cao, điều này cần thiết cho các tác vụ dự đoán ở mức điểm ảnh. Đồng thời, các phần chồng lấn lên nhau nhằm bảo toàn được thông tin có tính liên tục cục bộ. (ii) Kiến trúc mạng được xây dựng theo cấu trúc kim tự tháp có kích thước chiều không gian của đặc trưng giảm dần ở các tầng sâu hơn. Điều này giúp giảm đáng kể chi phí tính toán, đồng thời trích xuất được cả đặc trưng thô có độ phân giải cao và đặc trưng tinh có độ phân giải thấp. (iii) Thiết kế cơ chế chú ý hiệu quả nhằm giảm độ phức tạp và chi phí tính toán. (iv) Loại bỏ cơ chế mã hóa vị trí (position encoding - PE) trong Transformer truyền thống. Do trong quá trình huấn luyện, độ phân giải của PE có kích thước cố định, do đó nếu trong quá trình thực hiện kiểm tra hoặc suy luận, ảnh có độ phân giải khác với quá trình huấn luyện, các trọng số trong PE phải được nội suy, điều này gây giảm độ chính xác. Kiến trúc MiT đề xuất một mô-đun kết hợp giữa phép tích chập và tầng kết nối đầy đủ nhằm cung cấp thông tin vị trí cho mô hình, đồng thời có thể thích nghi với ảnh đầu vào có độ phân giải bất kỳ mà không ảnh hưởng đến hiệu năng của mô hình.

Tổng quan kiến trúc MiT được thể hiện trong Hình 3.2. Kiến trúc được chia thành 4 giai đoạn, mỗi giai đoạn là sự kết hợp của mô-đun tạo vector nhúng cho các phần của ảnh (overlapping patch embedding) và mô-đun mã hóa Transformer. Do được thiết kế theo cấu trúc kim tự tháp, độ phân giải của các đặc trưng giảm dần qua từng giai đoạn, nhờ đó tạo ra đặc trưng ở nhiều mức và nhiều tỉ lệ khác nhau (multi-scale feature), điều này phù hợp cho các tác vụ dự đoán mức điểm ảnh, giúp tăng độ chính xác khi phát hiện được đối tượng ở nhiều tỉ lệ khác nhau. Cụ thể, với một ảnh đầu vào, mô-đun này sẽ cho ra 4 đặc trưng $\{F_1, F_2, F_3, F_4\}$, có kích thước chiều không gian lần lượt bằng $\{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ kích thước ảnh đầu vào. 4 đặc

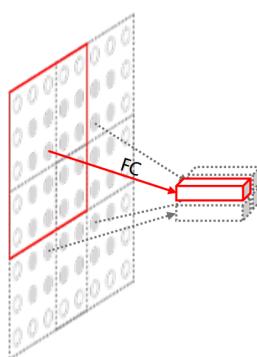
trưng này bao gồm F_1 là đặc trưng thô (đặc trưng mức thấp) có độ phân giải cao chứa thông tin chi tiết về hình dạng của polyp và F_2, F_3, F_4 là những đặc trưng tinh (đặc trưng ngữ nghĩa mức cao) có độ phân giải thấp, kết hợp những đặc trưng này thường tăng hiệu năng đáng kể của bài toán phân vùng ngữ nghĩa.



Hình 3.2: Kiến trúc Mix Transformer.

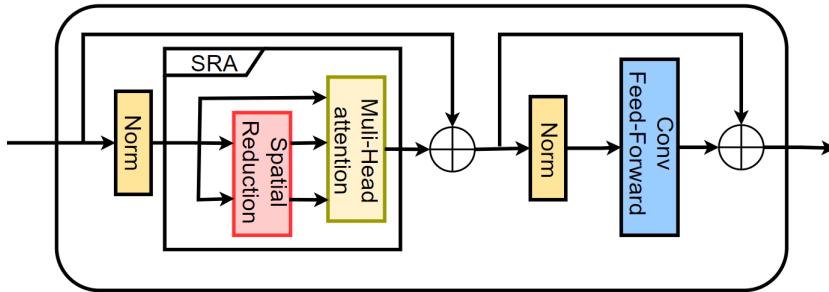
Chi tiết kiến trúc MiT:

- Tạo vector nhúng (overlapping patch embedding): trong kiến trúc ViT, ảnh đầu vào được chia thành các phần không chồng lấn nhau, do đó không bảo toàn được thông tin có tính liên tục cục bộ xung quanh mỗi phần. Để xử lý vấn đề này, MiT được thiết kế sao cho các phần ảnh liền kề chồng lấn nhau bởi một nửa diện tích (xem Hình 3.3). Nếu ảnh đầu vào có kích thước $H \times W \times C$, mỗi phần ảnh có kích thước $P \times P \times C$, dẫn đến kết quả đầu ra ảnh sẽ được chia thành $\frac{H \times W}{P^2}$ phần. Trong thực nghiệm, để chia ảnh thành các phần chồng lấn và thu được vector nhúng của mỗi phần, ta sử dụng một phép tích chập có sải bước (stride) S , bộ lọc có kích thước $2 \times S - 1$ và đệm (padding) có kích thước $S - 1$ nhằm tạo ra đặc trưng có kích thước bằng với kích thước khi không thực hiện chồng lấn các phần của ảnh.



Hình 3.3: Overlapping patch embedding.

- Mô-đun mã hóa Transformer ở giai đoạn i có Li khối xếp chồng lên nhau, mỗi khối là sự kết hợp giữa một tầng chú ý đa đầu (multi-head attention layer) và một tầng tích chập kết hợp với kết nối đầy đủ (convolutional feed-forward layer) (xem Hình 3.4).



Hình 3.4: Mô-đun Transformer trong MiT.

Trong mô-đun mã hóa Transformer của MiT, nút thắt trong quá trình tính toán nằm ở cơ chế chú ý, đồng thời MiT cần xử lý đặc trưng có độ phân giải cao hay phụ thuộc xa, nhóm tác giả đề xuất sử dụng mô-đun chú ý giảm chiều không gian (spatial-reduction attention - SRA) trong phần mã hóa nhằm giảm chi phí tính toán. Trong SRA, hai ma trận K và V được giảm chiều không gian với tỉ lệ R trước khi thực hiện tính ma trận chú ý.

Với cơ chế chú ý trong Transformer truyền thống:

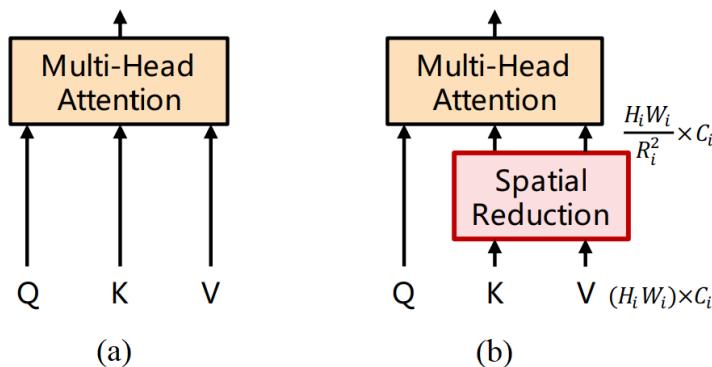
$$Q \in \mathbb{R}^{N \times E}, K \in \mathbb{R}^{N \times E}, V \in \mathbb{R}^{N \times E}$$

→ Khi đó, độ phức tạp tính toán sẽ là: $O(N^2 \times E)$

Với cơ chế chú ý giảm chiều không gian (SRA):

$$Q \in \mathbb{R}^{N \times E}, K \in \mathbb{R}^{\frac{N}{R} \times E}, V \in \mathbb{R}^{\frac{N}{R} \times E}$$

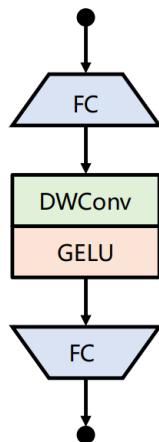
→ Khi đó, độ phức tạp tính toán sẽ là: $O\left(\frac{N^2}{R} \times E\right)$



Hình 3.5: Cơ chế tự chú ý trong: (a) Transformer truyền thống, (b) MiT.

- Do đã bỏ đi PE, nhóm tác giả SegFormer [7] đề xuất mô-đun Mix-FFN là sự

kết hợp giữa phép tích tích chập và tầng kết nối đầy đủ (xem Hình 3.6). Bằng kết quả thực nghiệm, nhóm tác giả cho thấy phép tích chập 3×3 cung cấp thông tin về vị trí cho Transformer một cách hiệu quả, đồng thời sử dụng phép tích chập tách biệt chiều sâu (depth-wise convolution) giúp giảm đáng kể số lượng tham số và cải thiện hiệu năng.



Hình 3.6: Mix-FFN.

Một số ưu điểm của kiến trúc MiT: (i) CNN truyền thống tạo ra trường tiếp nhận cục bộ (local receptive field), trong khi đó MiT luôn tạo ra trường tiếp nhận toàn cục (global receptive field), điều này phù hợp hơn với các bài toán phát hiện vật thể hay phân vùng ảnh. Tuy nhiên, nó cũng cho thấy những hạn chế trong việc tìm ra các đặc trưng chi tiết, đặc biệt trong xử lý ảnh y tế. (ii) MiT kế thừa những ưu điểm của cả CNN và Transformer, giúp tạo ra một bộ khung trích xuất đặc trưng linh hoạt cho các tác vụ thị giác máy tính, có thể được sử dụng trực tiếp thay thế cho các bộ khung CNN. (iii) Kiến trúc kim tự tháp và SRA giúp giảm chi phí tính toán, đồng thời giúp MiT linh hoạt hơn trong học đặc trưng ở nhiều mức và nhiều tỉ lệ khác nhau. (iv) Đưa ra đặc trưng có độ phân giải cao, điều này khá quan trọng cho các tác vụ dự đoán ở mức điểm ảnh.

MiT được thiết kế gồm 6 phiên bản ký hiệu từ B0 đến B5, các phiên bản này có kiến trúc giống nhau, chỉ khác nhau về siêu tham số (xem Hình 3.7). Trong đó:

- K: kích thước mỗi phần ảnh trong mô-đun overlapping patch embedding
- S: kích thước sải bước (stride) trong mô-đun overlapping patch embedding
- P: kích thước đệm (padding) trong mô-đun overlapping patch embedding
- C: số kênh (channel) của đầu ra
- L: số khôi mã hóa xếp chồng lên nhau
- R: tỉ lệ giảm chiều trong mô-đun chú ý giảm chiều không gian (SRA)

- N: số lượng đầu (head) trong tầng chú ý đa đầu (multi-head attention)
- E: tỉ lệ mở rộng của tầng kết nối đầy đủ (feed-forward)

	Output Size	Layer Name	Mix Transformer								
			B0	B1	B2	B3	B4	B5			
Stage 1	$\frac{H}{4} \times \frac{W}{4}$	Overlapping Patch Embedding	$K_1 = 7; S_1 = 4; P_1 = 3$								
			$C_1 = 32$	$C_1 = 64$							
Stage 2	$\frac{H}{8} \times \frac{W}{8}$	Overlapping Patch Embedding	$K_2 = 3; S_2 = 2; P_2 = 1$								
			$C_2 = 64$	$C_2 = 128$							
Stage 3	$\frac{H}{16} \times \frac{W}{16}$	Overlapping Patch Embedding	$K_3 = 3; S_3 = 2; P_3 = 1$								
			$C_3 = 160$	$C_3 = 320$							
Stage 4	$\frac{H}{32} \times \frac{W}{32}$	Overlapping Patch Embedding	$K_4 = 3; S_4 = 2; P_4 = 1$								
			$C_4 = 256$	$C_4 = 512$							
			Transformer Encoder	$R_1 = 8$	$R_1 = 8$	$R_1 = 8$	$R_1 = 8$	$R_1 = 8$	$R_1 = 8$		
				$N_1 = 1$	$N_1 = 1$	$N_1 = 1$	$N_1 = 1$	$N_1 = 1$	$N_1 = 1$		
			Transformer Encoder	$E_1 = 8$	$E_1 = 8$	$E_1 = 8$	$E_1 = 8$	$E_1 = 8$	$E_1 = 4$		
				$L_1 = 2$	$L_1 = 2$	$L_1 = 3$	$L_1 = 3$	$L_1 = 3$	$L_1 = 3$		
			Transformer Encoder	$R_2 = 4$	$R_2 = 4$	$R_2 = 4$	$R_2 = 4$	$R_2 = 4$	$R_2 = 4$		
				$N_2 = 2$	$N_2 = 2$	$N_2 = 2$	$N_2 = 2$	$N_2 = 2$	$N_2 = 2$		
			Transformer Encoder	$E_2 = 8$	$E_2 = 8$	$E_2 = 8$	$E_2 = 8$	$E_2 = 8$	$E_2 = 4$		
				$L_2 = 2$	$L_2 = 2$	$L_2 = 3$	$L_2 = 3$	$L_2 = 8$	$L_2 = 6$		
			Transformer Encoder	$R_3 = 2$	$R_3 = 2$	$R_3 = 2$	$R_3 = 2$	$R_3 = 2$	$R_3 = 2$		
				$N_3 = 5$	$N_3 = 5$	$N_3 = 5$	$N_3 = 5$	$N_3 = 5$	$N_3 = 5$		
			Transformer Encoder	$E_3 = 4$	$E_3 = 4$	$E_3 = 4$	$E_3 = 4$	$E_3 = 4$	$E_3 = 4$		
				$L_3 = 2$	$L_3 = 2$	$L_3 = 6$	$L_3 = 18$	$L_3 = 27$	$L_3 = 40$		
			Transformer Encoder	$R_4 = 1$	$R_4 = 1$	$R_4 = 1$	$R_4 = 1$	$R_4 = 1$	$R_4 = 1$		
				$N_4 = 8$	$N_4 = 8$	$N_4 = 8$	$N_4 = 8$	$N_4 = 8$	$N_4 = 8$		
			Transformer Encoder	$E_4 = 4$	$E_4 = 4$	$E_4 = 4$	$E_4 = 4$	$E_4 = 4$	$E_4 = 4$		
				$L_4 = 2$	$L_4 = 2$	$L_4 = 3$	$L_4 = 3$	$L_4 = 3$	$L_4 = 3$		

Hình 3.7: Các phiên bản của kiến trúc Mix Transformer [7].

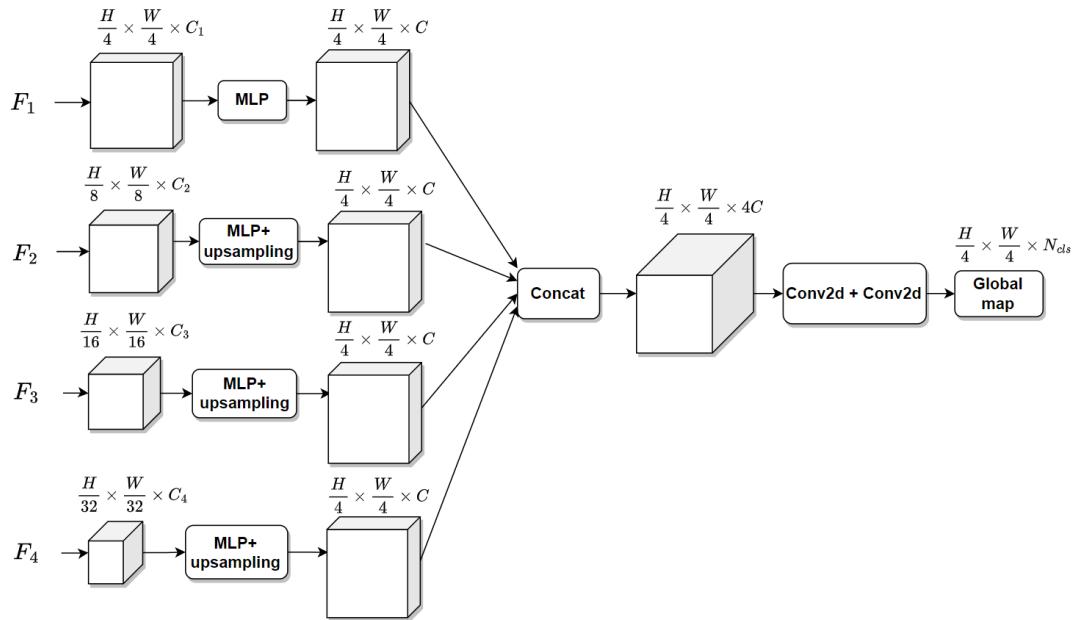
Để cân bằng giữa độ chính xác và thời gian suy luận, trong kiến trúc đề xuất sử dụng Mix Transformer phiên bản MiT-B3, phiên bản này có 44.725.696 tham số và số lượng phép tính dấu phẩy động là 17.884 GFLOPs. Trọng số khởi tạo của mạng sử dụng trọng số được tiền huấn luyện trên bộ dữ liệu ImageNet.

3.3 Mô-đun giải mã

3.3.1 Mô-đun giải mã đề xuất

Dựa trên mô-đun giải mã MLP được giới thiệu trong bài báo SegFormer [7]. Chi tiết kiến trúc của mô-đun này được biểu diễn ở Hình 3.8. Trong bài báo đã chứng minh đây là một kiến trúc mạng nhẹ (lightweight network) được thiết kế đơn giản nhằm tận dụng được ưu điểm từ các đặc trưng cho ra bởi kiến trúc MiT là có trường tiếp nhận cục bộ lớn nhờ cấu trúc phân tầng và tạo ra đặc trưng toàn cục. Cụ thể, với các đặc trưng cho ra bởi MiT, cơ chế chú ý tại các tầng thấp hơn có xu hướng duy trì tính cục bộ, trong khi tại các tầng cao hơn có xu hướng phi cục bộ cao. Với bài toán phân vùng ngữ nghĩa, việc duy trì trường tiếp nhận lớn để trích xuất hiệu quả thông tin ngữ cảnh là một vấn đề trọng tâm. Mô-đun này tích hợp thông tin từ

các tầng khác nhau, do đó tích hợp được cả thông tin chú ý cục bộ (local attention) và thông tin chú ý toàn cục (global attention) để tạo ra các biểu diễn đặc trưng mạnh mẽ.



Hình 3.8: Mô-đun giải mã trong bài báo SegFormer [7].

Tuy nhiên, do tồn tại sự tương quan thấp giữa các đặc trưng tại các tầng khác nhau của cấu trúc kim tự tháp, tích hợp các đặc trưng này một cách song song như được thiết kế ban đầu trong bài báo SegFormer [7] có thể tạo ra lỗ hổng thông tin. Do đó trong đồ án này đề xuất tích hợp thêm mô-đun CBAM nhằm tìm ra các đặc trưng quan trọng cần chú ý trước khi thực hiện tích hợp. Mô-đun CBAM là cơ chế chú ý thuần CNN, đây cũng là một kiến trúc mạng nhẹ, chỉ bao gồm các lớp tích chập, có số lượng tham số nhỏ, dễ dàng tích hợp vào kiến trúc mạng cơ sở. Mô-đun này có thể đánh trọng số thể hiện độ quan trọng của đặc trưng theo cả chiều không gian và chiều sâu (được trình bày chi tiết trong **Mục 3.3.2**).

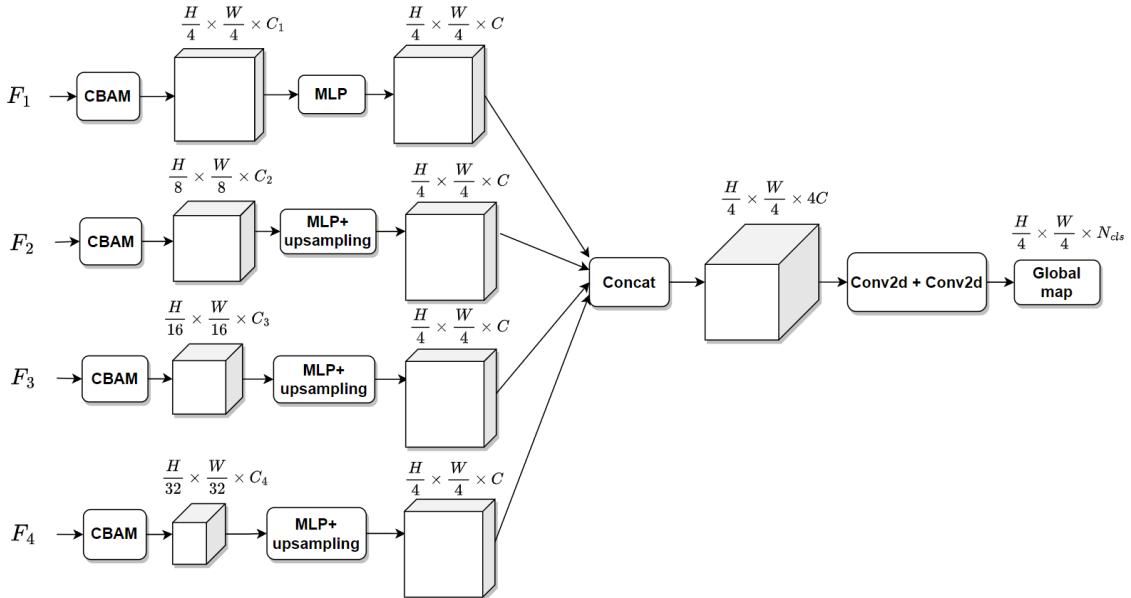
Chi tiết kiến trúc mô-đun giải mã đề xuất được thể hiện ở **Hình 3.9**. Mô-đun giải mã này bao gồm 4 bước thực hiện chính:

Bước 1: các đặc trưng ở nhiều tỉ lệ khác nhau F_i được cho ra bởi mô-đun mã hóa MiT lần lượt được đưa qua mô-đun CBAM để tinh chỉnh đặc trưng và perceptron đa tầng để đồng nhất số chiều sâu. Trong đồ án này, các đặc trưng được đồng nhất về chiều sâu có độ lớn 256.

Bước 2: các đặc trưng được chỉnh lại kích thước sử dụng toán tử nội suy song tuyến (bilinear interpolation) sao cho kích thước chiều không gian bằng $\frac{1}{4}$ so với kích thước ảnh huấn luyện, tiếp đó 4 đặc trưng được nối với nhau theo chiều sâu.

Bước 3: sử dụng một tầng tích chập để hợp nhất đặc trưng vừa được nối với nhau.

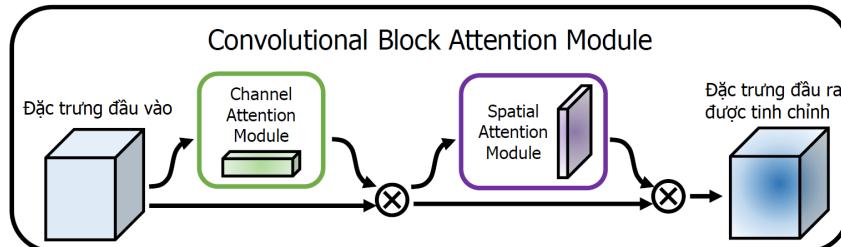
Bước 4: đưa qua một tầng tích chập khác để dự đoán ra bản đồ đặc trưng toàn cục có kích thước $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ (N_{cls} là số lớp cần phân vùng).



Hình 3.9: Mô-đun giải mã đề xuất.

3.3.2 Convolutional Block Attention Module (CBAM)

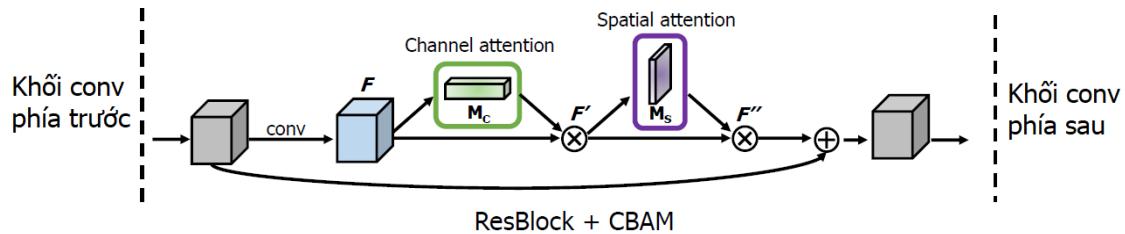
Kiến trúc CBAM [21] được thiết kế lấy cảm hứng từ cơ chế chú ý, nhằm tinh chỉnh đặc trưng và cải thiện khả năng biểu diễn thông tin tại vùng chứa đối tượng cần chú ý (hay đối tượng đích). CBAM trích xuất thông tin từ đặc trưng bằng cách pha trộn thông tin từ cả miền không gian và miền chiều sâu. Kiến trúc CBAM là sự kết hợp của hai mô-đun: chú ý theo kênh (channel attention module - CAM) và chú ý theo không gian (spatial attention module - SAM). Tổng quan kiến trúc CBAM được minh họa trong Hình 3.10:



Hình 3.10: Tổng quan kiến trúc CBAM.

Khác với cơ chế chú ý của Transformer, CBAM là một mô-đun thuần CNN, chỉ bao gồm các lớp tích chập. CBAM được thiết kế sao cho số lượng tham số và số phép tính cần tính nhỏ (light-weight network) mà vẫn mang lại hiệu năng cao, đồng

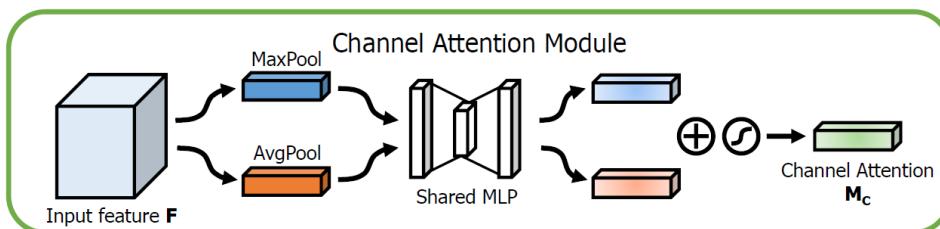
thời có thể dễ dàng tích hợp vào kiến trúc mạng cơ sở nhằm tinh chỉnh đặc trưng được cho ra bởi tầng tích chập bất kỳ (xem Hình 3.11).



Hình 3.11: Minh họa cách tích hợp CBAM vào kiến trúc cơ sở.

a, Mô-đun chú ý theo kênh (CAM)

Mô-đun này được xây dựng nhằm mục đích khai thác mối liên hệ giữa các kênh của đặc trưng. Mỗi kênh này có nhiệm vụ lưu trữ các đặc trưng khác nhau từ ảnh đầu vào, CAM tập trung tìm ra đặc trưng nào có ý nghĩa bằng cách đánh trọng số cho mỗi kênh.



Hình 3.12: Mô-đun chú ý theo kênh [21].

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W(F_{avg}^c) + W(F_{max}^c)) \end{aligned} \quad (3.1)$$

Chi tiết cơ chế hoạt động của mô-đun được minh họa trong Hình 3.12 và tổng hợp trong công thức (3.1), bao gồm 3 bước chính:

Bước 1: với mỗi kênh của đặc trưng đầu vào, áp dụng đồng thời hai toán tử gộp trung bình toàn cục (Global Average Pooling - GAP) và gộp cực đại toàn cục (Global Max Pooling - GMP), dẫn đến sau khi thực hiện mỗi toán tử, mỗi kênh sẽ bị nén thành một số vô hướng. Nếu đặc trưng đầu vào có C kênh, sau khi áp dụng hai toán tử này sẽ thu được hai vector, mỗi vector có kích thước $[1 \times 1 \times C]$. GAP được sử dụng nhằm tạo ra thông tin kết hợp từ chiều không gian, trong khi đó sử dụng GMP nhằm bảo toàn thông tin kết cấu như cạnh hay biên của đối tượng.

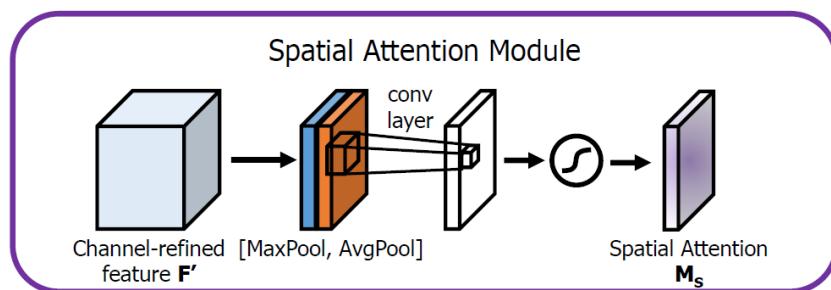
Bước 2: hai vector được tạo bởi GAP và GMP được đưa qua một mạng perceptron đa tầng có chia sẻ trọng số, kết quả đầu ra cũng tạo được hai vector có

kích thước [1x1xC].

Bước 3: hai vector vừa tạo ra được cộng với nhau tương ứng từng phần tử, sau đó đưa qua hàm kích hoạt Sigmoid nhằm đưa giá trị về đoạn [0, 1]. Cuối cùng ta thu được một vector có kích thước [1x1xC] với ý nghĩa lưu giữ trọng số độ quan trọng mỗi kênh của đặc trưng.

b, Mô-đun chú ý theo không gian (SAM)

Mô-đun này tập trung tìm ra vị trí nào trên đặc trưng chứa thông tin quan trọng cần chú ý bằng cách đánh trọng số cho mỗi điểm ảnh trên bản đồ đặc trưng.



Hình 3.13: Mô-đun chú ý theo không gian [21].

$$\begin{aligned} M_s(F) &= \sigma(f^{7 \times 7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7 \times 7}([F_{avg}^s; F_{max}^s])) \end{aligned} \quad (3.2)$$

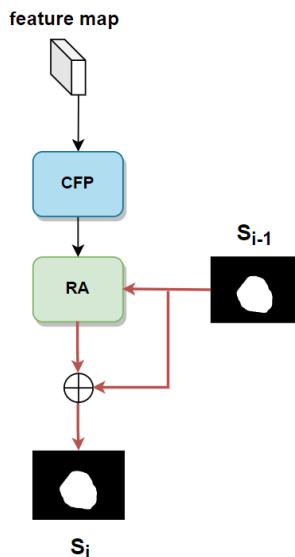
Chi tiết cơ chế hoạt động của mô-đun được minh họa trong **Hình 3.13** và tổng hợp trong công thức (3.2), bao gồm 2 bước chính:

Bước 1: đặc trưng đầu vào đồng thời được đưa qua toán tử gộp cực đại (Max Pooling - MP) và gộp trung bình (Average Pooling - AP), hai toán tử này được thực hiện theo chiều sâu của đặc trưng. Do đó, nếu đặc trưng đầu vào có kích thước [HxWxC] (C là số kênh), sau mỗi toán tử trên sẽ thu được đặc trưng đầu ra có kích thước [HxWx1].

Bước 2: hai đặc trưng đầu ra sau khi thực hiện MP và AP được nối với nhau theo chiều sâu, thu được đặc trưng mới có kích thước [HxWx2]. Tiếp theo đó đưa qua một tầng tích chập với kích thước kernel=7 và padding=3 nhằm tạo ra đặc trưng đầu ra có kích thước chiều không gian không đổi và kích thước chiều sâu giảm về 1. Cuối cùng, đặc trưng đầu ra này được đưa qua hàm kích hoạt Sigmoid nhằm đưa giá trị trọng số về đoạn [0, 1], thu được đặc trưng có kích thước [HxWx1] với ý nghĩa lưu giữ trọng số độ quan trọng của từng điểm ảnh.

3.4 Mô-đun tinh chỉnh đặc trưng

Trong quá trình nội soi đại tràng, bước đầu tiên bác sĩ nội soi sẽ xác định vị trí tương đối của polyp, sau đó xem xét cẩn thận vùng niêm mạc xung quanh để đánh nhận polyp một cách chính xác nhất.



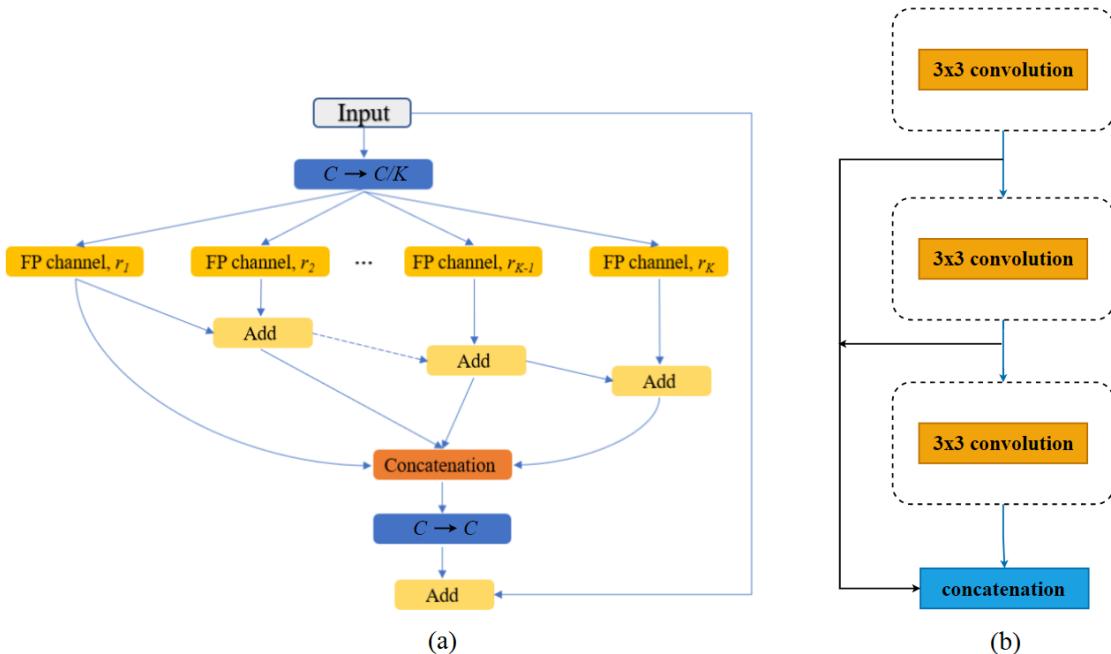
Hình 3.14: Tổng quan mô-đun tinh chỉnh đặc trưng.

Bản đồ đặc trưng toàn cục cho ra bởi mô-đun giải mã chỉ xác định được vị trí tương đối của polyp mà không mang nhiều thông tin cấu trúc chi tiết. Do đó, dựa theo phân tích của quá trình nội soi thực tế, trong đồ án này sử dụng mô-đun tinh chỉnh đặc trưng nhằm tinh chỉnh bản đồ đặc trưng toàn cục để đạt được độ chính xác và cho ra kết quả dự đoán hoàn thiện hơn. Mô-đun tinh chỉnh được thiết kế bao gồm hai mô-đun con: đặc trưng kim tự tháp theo kênh (channel-wise feature pyramid - CFP) và chú ý ngược (reverse attention - RA) (xem Hình 3.14).

3.4.1 Trích xuất đặc trưng kim tự tháp theo kênh (CFP)

Kiến trúc trích xuất đặc trưng kim tự tháp (Feature Pyramid) được sử dụng rộng rãi trong các mô hình học sâu, đặc biệt trong tác vụ thị giác máy tính thực hiện dự đoán ở mức điểm ảnh bởi khả năng biểu diễn đặc trưng ở nhiều tỉ lệ khác nhau. Trong bài báo CFPNet [22] đã đề xuất mô-đun trích xuất đặc trưng kim tự tháp theo kênh (Channel-wise Feature Pyramid) lấy ý tưởng từ những ưu điểm của mạng Inception và tích chập dần nở. Việc học song song và áp dụng tích chập dần nở giúp mạng có thể học được nhiều chi tiết hơn, lấy được nhiều đặc trưng ở nhiều tỉ lệ khác nhau, trích xuất đặc trưng có trường tiếp nhận cục bộ lớn trong khi bảo toàn số lượng tham số cần học. Đây là một mô-đun cân bằng được cả 3 yếu tố: hiệu năng, kích thước và tốc độ. Kết quả thực nghiệm cho thấy CFP đạt được cải tiến đáng kể trong tác vụ thị giác máy tính đối với ảnh thông thường và ảnh y tế.

Tổng quan kiến trúc mô-đun CFP:



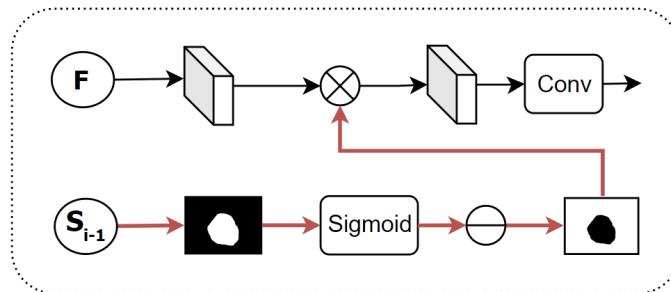
Hình 3.15: (a) Mô-đun CFP, (b) Mô-đun FP.

Với C là số chiều sâu của đặc trưng đầu vào, ta sẽ áp dụng một tầng tích chập sao cho kết quả đặc trưng đầu ra sẽ có chiều sâu là C/K . Đặc trưng này sẽ được song song đưa qua K kênh, mỗi kênh là một mô-đun FP có tỉ lệ tích chập dãn nở r khac nhau. Thông thường K được chọn bằng 4 và tỉ lệ tích chập dãn nở ở mỗi kênh $\{r_1, r_2, r_3, r_4\} = \{1, 2, 4, 8\}$. Ngoài ra, mô-đun FP chỉ đơn thuần bao gồm 3 tầng tích chập dãn nở xếp chồng lên nhau, đầu ra của cả 3 tầng được nối với nhau theo chiều sâu để tạo kết quả cuối cùng (xem Hình 3.15 (b)). Trong CFP, kể từ đầu ra tại mô-đun FP thứ 2 sẽ được kết hợp với đầu ra của mô-đun FP tại bước trước thông qua toán tử cộng tương ứng từng phần tử. Sau khi hoàn thành K mô-đun FP, các kết quả đầu ra của K mô-đun FP này sẽ được nối với nhau theo chiều sâu (xem Hình 3.15 (a)).

3.4.2 Mô-đun chú ý ngược (RA) và kết nối phần dư

Phương pháp này được trình bày chi tiết trong bài báo Reverse Attention for Salient Object Detection [23], được tích hợp vào kiến trúc mạng cho bài toán phát hiện đối tượng nổi bật trong ảnh (Salient Object Detection- SOD). Với bài toán phân vùng ngữ nghĩa ảnh, mô hình cần phân vùng tất cả đối tượng trong ảnh, trong khi đó, SOD nhằm mục đích tìm vị trí và phân vùng đối tượng nổi bật hay thu hút nhất trong ảnh. RA được thiết kế nhằm tinh chỉnh bản đồ đặc trưng, giúp kiến trúc mạng tập trung học biểu diễn các phần thông tin chưa được khám phá, đặc biệt là biên của đối tượng.

Tổng quan mô-đun RA được minh họa trong Hình 3.16. Đầu vào của mô-đun RA gồm hai phần: (i) bản đồ đặc trưng F: chứa đặc trưng của đối tượng và (ii) đầu ra biên S_{i-1} : là kết quả dự đoán của mô hình tại bước trước đó, kết quả này đang cần được tinh chỉnh.



Hình 3.16: Mô-đun chú ý ngược (RA).

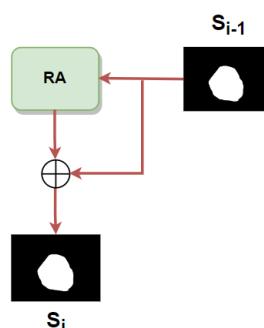
Cơ chế hoạt động của mô-đun RA:

- Tạo mặt nạ bằng cách xóa đi vùng dự đoán của S_{i-1} (hay vùng đối tượng đã được phát hiện từ đầu ra của bước trước) thông qua công thức:

$$R = 1 - \text{Sigmoid}(S_{i-1})$$

- Mặt nạ vừa tạo ra được nhân với bản đồ đặc trưng F bằng toán tử nhân tương ứng từng phần tử nhằm đóng băng thông tin đã được học trước đó, buộc mạng tập trung học để tìm ra thông tin tại vùng đối tượng chưa được phát hiện, giúp khám phá ra các phần đối tượng bị thiếu. Nhờ thiết kế này mà mạng có thể học được tốt chi tiết phần biên của đối tượng.

Đầu ra của mô-đun RA là đặc trưng chứa thông tin còn thiếu nhằm bổ sung cho đầu ra dự đoán của bước trước. Do đó ta cần một kết nối phần dư kết hợp đầu ra của mô-đun RA và đầu ra của bước trước thông qua toán tử cộng tương ứng từng phần tử để thu được kết quả cuối cùng đã được tinh chỉnh (xem Hình 3.17).



Hình 3.17: Kết nối phần dư bổ sung thông tin trích xuất được từ mô-đun RA cho đầu ra dự đoán của bước trước S_{i-1} .

Trong PraNet [2] và CaraNet [20] cũng đề xuất mô-đun tinh chỉnh đặc trưng, nhưng chỉ thực hiện tinh chỉnh tại 3 đặc trưng ở mức cao nhất cho ra bởi mô-đun mã hóa. Tuy nhiên, em nhận thấy tại các đặc trưng mức thấp đầu tiên của mô-đun mã hóa thường giàu thông tin chi tiết như thông tin kết cấu, màu sắc hay thông tin về cạnh. Do đó trong đồ án này đề xuất thực hiện tinh chỉnh tại cả 4 đặc trưng cho ra bởi mô-đun mã hóa.

3.5 Định nghĩa hàm mất mát

Hàm mất mát là sự kết hợp giữa Binary Cross-Entropy (BCE) và mean Intersection over Union (mIoU). Trong đó, BCE nhằm tính toán sự mất mát cục bộ (hay ở mức điểm ảnh), đo sự khác nhau giữa phân phối dự đoán của mô hình và phân phối thực; mIoU nhằm tính toán sự mất mát toàn cục, đo sự khác nhau giữa nhãn phân vùng dự đoán và nhãn thực. Đồng thời có sử dụng trọng số cho từng điểm ảnh và trọng số này tập trung hơn vào các điểm ảnh là biên của polyp.

$$L = \frac{L_{IoU}^w + L_{BCE}^w}{2} \quad (3.3)$$

$$L_{IoU}^w = 1 - \frac{\sum_{i=1}^H \sum_{j=1}^W (y_{ij} \times \hat{y}_{ij} \times w_{ij}) + 1}{\sum_{i=1}^H \sum_{j=1}^W ((y_{ij} + \hat{y}_{ij}) \times w_{ij} - y_{ij} \times \hat{y}_{ij} \times w_{ij}) + 1} \quad (3.4)$$

$$L_{BCE}^w = -\frac{\sum_{i=1}^H \sum_{j=1}^W w_{ij} (y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij}))}{\sum_{i=1}^H \sum_{j=1}^W w_{ij}} \quad (3.5)$$

$$w_{ij} = 1 + \left| \frac{\sum_{m,n \in A_{ij}} y_{mn}}{\sum_{m,n \in A_{ij}} 1} - y_{ij} \right| \quad (3.6)$$

Trong đó:

- y_{ij} là giá trị nhãn thực (ground truth) tại điểm ảnh (i, j)
- \hat{y}_{ij} là giá trị mô hình dự đoán tại điểm ảnh (i, j)
- A_{ij} là vùng diện tích lân cận xung quanh điểm ảnh (i, j)
- w_{ij} ước tính độ quan trọng của điểm ảnh (i, j) bằng cách đo độ khác biệt giữa giá trị nhãn thực và giá trị các điểm ảnh lân cận xung quanh (i, j) . w_{ij} càng lớn thể hiện điểm ảnh này có sự khác biệt đáng kể với các điểm ảnh xung quanh, tức điểm ảnh này là đường biên hoặc nằm gần biên của polyp.

Hàm mất mát được áp dụng lên bản đồ đặc trưng toàn cục S_g và cả 4 đầu ra của mô-đun tinh chỉnh đặc trưng $\{S_1, S_2, S_3, S_4\}$ (xem Hình 3.1). Trước khi tính toán

hàm mất mát, các đầu ra này được chỉnh lại kích thước sao cho có cùng kích thước với nhãn thực thông qua phép nội suy song tuyến. Do đó, hàm mất mát tổng thể của mô hình sẽ là:

$$L_{total} = L(G, S_g) + \sum_{i=1}^4 L(G, S_i) \quad (3.7)$$

Kết chương: trong chương này đã trình bày chi tiết 3 mô-đun của mô hình đề xuất. Mô-đun mã hóa sử dụng Mix-Transformer; mô-đun giải mã tích hợp song song các đặc trưng cho ra bởi mô-đun mã hóa với sự hỗ trợ của CBAM nhằm tạo bản đồ đặc trưng toàn cục; mô-đun tinh chỉnh đặc trưng là sự kết hợp giữa kiến trúc CFP và RA nhằm tinh chỉnh bản đồ đặc trưng toàn cục, tập trung hơn vào các phần thông tin chưa được khám phá, đặc biệt là biên của polyp. Chương này cũng định nghĩa hàm mất mát cần tối ưu, hàm mất mát này là sự kết hợp giữa IoU và BCE, trong đó có sử dụng trọng số cho từng điểm ảnh nhằm mục đích tập trung hơn vào các điểm ảnh là biên của polyp.

CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM

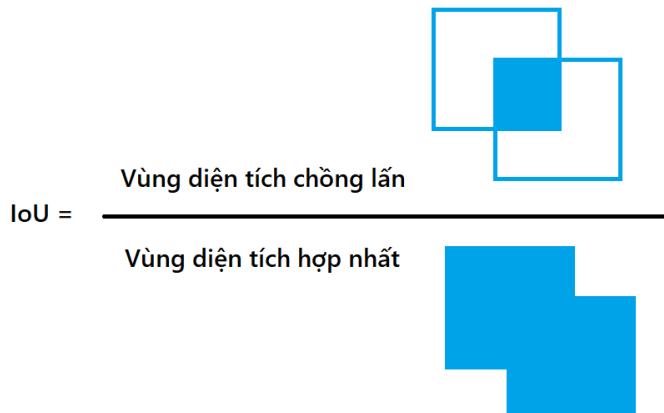
Trong chương 3 đã trình bày chi tiết về mô hình đề xuất. Chương 4 này sẽ trình bày các độ đo hiệu năng của mô hình, bộ dữ liệu sử dụng, phương pháp thực nghiệm nhằm so sánh, đánh giá mô hình đề xuất với các mô hình khác và chứng minh sự hiệu quả của các mô-đun trong mô hình đề xuất.

4.1 Độ đo đánh giá

Để đánh giá hiệu năng của một mô hình phân vùng ảnh, thông thường sẽ sử dụng một số độ đo phổ biến như: Intersection over Union (Jaccard Index) hoặc Dice coefficient (F1 score).

4.1.1 Intersection over Union (IoU, Jaccard index)

Intersection over Union (IoU), cũng được biết đến là độ đo Jaccard index, đây là một độ đo được sử dụng phổ biến và hiệu quả trong phân vùng ngữ nghĩa. IoU được tính bằng cách lấy diện tích vùng chồng lấn của nhãn dự đoán và nhãn thực chia cho vùng diện tích hợp nhất của cả hai nhãn (xem hình minh họa 4.1).



Hình 4.1: Minh họa hình học độ đo IoU.

IoU có độ lớn nằm trong đoạn $[0, 1]$, với 0 có nghĩa nhãn dự đoán và nhãn thực không có vùng chồng lấn, 1 có nghĩa là nhãn dự đoán và nhãn thực chồng lấn hoàn toàn.

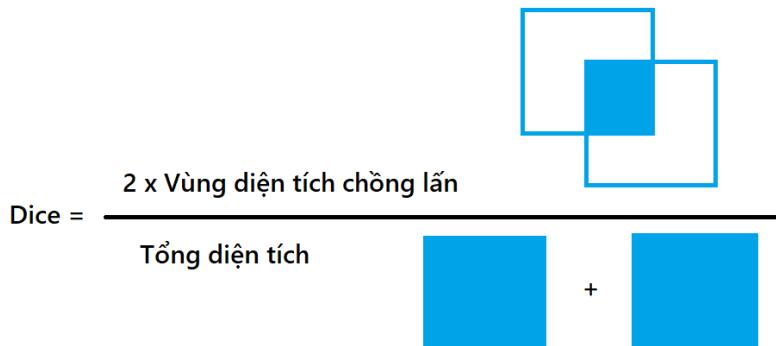
IoU được tính toán cho mỗi phân lớp riêng biệt, nếu nhãn dự đoán có nhiều lớp, ta sẽ tính giá trị trung bình của IoU theo công thức (4.1):

$$meanIoU = \frac{1}{c} \sum_c IoU_c \quad (4.1)$$

Trong đó c là số lượng các phân lớp

4.1.2 Dice coefficient

Dice coefficient là độ đo có liên hệ chặt chẽ và tương quan dương với IoU, độ đo này được tính bằng cách lấy 2 lần diện tích chồng lấn giữa nhãn dự đoán và nhãn thực chia cho tổng diện tích của hai nhãn (xem hình minh họa 4.2).



Hình 4.2: Minh họa hình học độ đo Dice.

Dice coefficient cũng có độ lớn nằm trong đoạn [0, 1], với 0 có nghĩa nhãn dự đoán và nhãn thực không có vùng chồng lấn, 1 có nghĩa là nhãn dự đoán và nhãn thực chồng lấn hoàn toàn.

Dice và IoU có công thức và cách tính toán gần như tương tự nhau, tuy nhiên sự khác biệt của hai độ đo này thể hiện khi tính trung bình trên tổng thể bộ dữ liệu. IoU có xu hướng phạt nặng hơn với các điểm ảnh bị phân loại sai. Mối liên hệ giữa IoU và Dice được thể hiện trong công thức (4.2):

$$IoU = \frac{Dice}{2 - Dice} \quad (4.2)$$

Xét về khía cạnh ma trận nhầm lẫn, các độ đo có thể được biểu diễn lại như trong công thức (4.3) và (4.4):

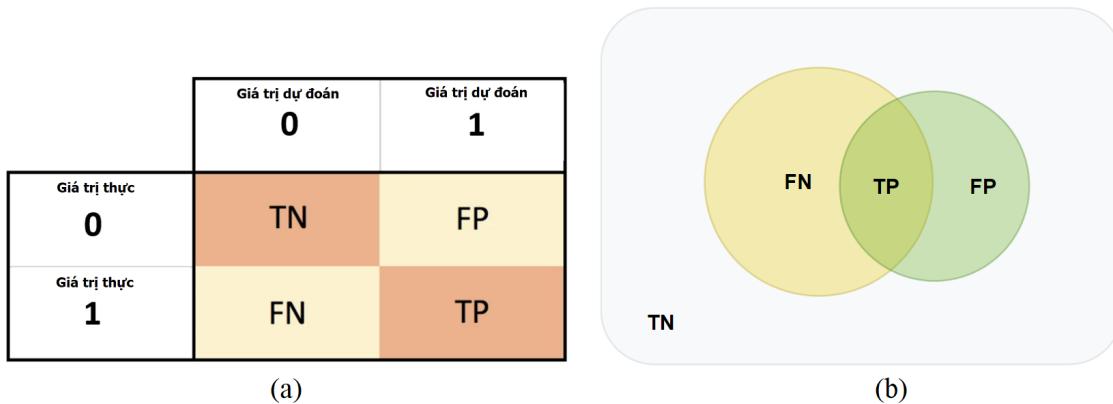
$$IoU = \frac{TP}{TP + FN + FP} \quad (4.3)$$

$$Dice = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (4.4)$$

Trong đó:

- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

Giả sử trong trường hợp phân vùng có hai nhãn lớp: 0 - nền, 1 - vật thể, ma trận nhầm lẫn có thể được minh họa như trong **Hình 4.3**.



Hình 4.3: (a) Ma trận nhầm lẫn và (b) Minh họa hình học của ma trận nhầm lẫn trong trường hợp phân vùng có hai nhãn lớp.

4.2 Dữ liệu

Bộ dữ liệu huấn luyện và kiểm thử mô hình bao gồm 5 bộ dữ liệu với các thông số được mô tả trong **Bảng 4.1**: Kvasir [24], CVC-Clinic DB [25], CVC-Colon DB [26], EndoScene [27] và ETIS-Larib Polyp DB [28]. Đây là những bộ dữ liệu mở chứa ảnh nội soi polyp đại tràng, được sử dụng phổ biến trong cộng đồng nghiên cứu:

- Kvasir: bộ dữ liệu được thu thập sử dụng thiết bị nội soi tại Vestre Viken Health Trust (VV), Na Uy. Ảnh nội soi được gán nhãn polyp bởi các chuyên gia y tế của VV và Cancer Registry, Na Uy. Bộ dữ liệu chứa 1000 ảnh và nhãn tương ứng với độ phân giải khác nhau, từ 332x487 đến 1920x1072 pixel.
- CVC-Clinic DB: bộ dữ liệu được cắt ra từ video nội soi được thu thập từ Hospital Clinic, Barcelona, Tây Ban Nha. Bộ dữ liệu bao gồm 612 ảnh có độ phân giải 384x288 pixel.
- CVC-Colon DB: bao gồm 380 ảnh có độ phân giải 574x500 pixel.
- EndoScene: bao gồm 612 ảnh từ tập dữ liệu CVC-Clinic DB và 300 ảnh từ tập CVC-300, cho mục đích huấn luyện mô hình. Tập dữ liệu kiểm tra bao gồm 60 ảnh có độ phân giải 574x500, trong một số bài báo được gọi với tên EndoScene hay CVC-T.
- ETIS-Larib Polyp DB: bộ dữ liệu được thu nhận từ phòng thí nghiệm ETIS, đại học Cergy-Pontoise, Pháp; chứa 196 ảnh có độ phân giải cao 1225x966 pixel.

Bảng 4.1: Thống kê thông số các bộ dữ liệu

Bộ dữ liệu		Số lượng ảnh	Độ phân giải
Kvasir		1000	332x487 - 1920x1072
CVC-ClinicDB		612	384 x 288
CVC-ColonDB		380	574 x 500
ETIS-Larib Polyp DB		196	1225 x 966
EndoScene	Đào tạo	CVC-ClinicDB	612
		CVC-300	300
	Kiểm thử	CVC-T	60
			574 x 500

4.3 Phương pháp tiến hành thực nghiệm

Mã nguồn được thực thi dựa trên framework PyTorch, quá trình huấn luyện mô hình sử dụng Google Colab. Thông số phần cứng được thống kê trong **Bảng 4.2.**

Bảng 4.2: Thống kê thông số phần cứng trên Google Colab

GPU	NVIDIA Tesla T4
Bộ nhớ trong	16 GB
Số phép tính dấu phẩy động / giây	8.1 TFLOPS

Số lượng tham số, số phép tính dấu phẩy động khi thực hiện suy luận với mỗi ảnh (GFLOPs) và kích thước lưu trữ của mô hình đề xuất được thống kê tại **Bảng 4.3.**

Bảng 4.3: Thống kê thông số của mô hình đề xuất

Số lượng tham số	45.687.829
GFLOPs	20.719
Kích thước lưu trữ mô hình	174 MB

Chiến lược huấn luyện mô hình: mô hình được huấn luyện trong 20 epoch với kích thước batch là 8. Ảnh đầu vào được đưa về kích thước 352x352, sau đó thực hiện chuẩn hóa ảnh bằng cách trừ cho giá trị trung bình và chia cho giá trị độ lệch chuẩn tương ứng trên ba kênh R, G, B như trong công thức (4.5). Các giá trị trung bình và độ lệch chuẩn này được lấy theo giá trị thống kê trên bộ dữ liệu có kích thước lớn ImageNet:

$$\text{Giá trị trung bình} = [0.485, 0.456, 0.406]$$

$$\text{Giá trị độ lệch chuẩn} = [0.229, 0.224, 0.225]$$

$$i_{out} = \frac{i_{in} - m_c}{s_c} \quad (4.5)$$

Trong đó:

- i_{in} là giá trị điểm ảnh đầu vào
- m_c là giá trị trung bình tại kênh c
- s_c là giá trị độ lệch chuẩn tại kênh c
- i_{out} là giá trị điểm ảnh sau khi chuẩn hóa

Ngoài trừ trọng số khởi tạo của mô-đun mã hóa MiT sử dụng trọng số được tiền huấn luyện trên bộ dữ liệu ImageNet, hai mô-đun còn lại là mô-đun giải mã và mô-đun tinh chỉnh đặc trưng sử dụng phương pháp khởi tạo He normal. Ngoài ra, sử dụng chiến lược huấn luyện ở nhiều tỉ lệ ảnh khác nhau. Trong đó, ảnh đầu vào được sử dụng ở các tỉ lệ (0.75, 1, 1.25). Phương pháp tối ưu sử dụng thuật toán Adam với tốc độ học (learning rate) khởi tạo bằng 1e-4. Đồng thời, giá trị tốc độ học giảm dần qua mỗi epoch theo hàm số:

$$\begin{cases} lr \times 0.1, \text{ nếu } e \geq 10 \text{ và } e \mod 5 \\ lr, \text{ các trường hợp còn lại} \end{cases}, e \text{ là chỉ số epoch hiện tại}$$

Trong quá trình lan truyền ngược cập nhật trọng số, sử dụng kỹ thuật gradient clipping nhằm mục đích giới hạn giá trị gradient sao cho luôn nằm trong đoạn [-0.5, 0.5].

Phương pháp thực nghiệm: để đánh giá hiệu năng của mô hình đề xuất, trong đồ án thực hiện huấn luyện mô hình sử dụng dữ liệu là sự kết hợp của hai tập Kvasir và CVC-ClinicDB, với mỗi tập sẽ trích ra 80% dữ liệu cho tập huấn luyện và 10% dữ liệu cho tập phát triển. Sau đó, sẽ sử dụng 10% dữ liệu còn lại của hai tập Kvasir và CVC-ClinicDB để đánh giá khả năng học của mô hình. Cụ thể, tập dữ liệu huấn luyện sẽ bao gồm 1305 ảnh, tập phát triển bao gồm 145 ảnh và tập kiểm tra đánh giá khả năng học của mô hình bao gồm 100 ảnh của Kvasir và 62 ảnh của CVC-ClinicDB. Mặt khác, sử dụng toàn bộ các tập CVC-ColonDB, ETIS và tập kiểm tra của EndoScene (CVC-T) - đây là các tập dữ liệu mà mô hình không sử dụng trong quá trình huấn luyện, do đó sử dụng các tập này nhằm đánh giá khả năng tổng quát hóa của mô hình. Việc huấn luyện và kiểm định mô hình được thực hiện 5 lần, kết quả thí nghiệm thu được bằng cách tính trung bình các kết quả của 5 lần chạy.

Kết quả thực nghiệm được so sánh với các mô hình tiên tiến nhất (SOTA) cho bài toán phân vùng polyp trên ảnh nội soi, bao gồm 9 mô hình: U-Net [15], UNet++ [16], SFA [29], PraNet [2], CaraNet [20], HarDNet-MSEG [19], TransUNet [3], TransFuse [4] và Polyp-PVT [5]. Trong đó, U-Net, UNet++, SFA, PraNet, CaraNet,

HarDNet-MSEG là các phương pháp chỉ sử dụng mạng nơron tích chập; TransUNet, TransFuse, Polyp-PVT là các phương pháp kết hợp giữa mạng nơron tích chập và Transformer. Độ đo được sử dụng để đánh giá và so sánh là hard mean Dice (mDice) và hard mean IoU (mIoU) với ngưỡng nhị phân được chọn là 0.5.

4.4 So sánh kết quả thực nghiệm các mô hình

Bảng 4.4 trình bày kết quả đánh giá các mô hình trên 5 tập dữ liệu kiểm tra. Từ bảng này, có thể nhận thấy rằng độ chính xác của mô hình đề xuất cao hơn các mô hình khác trên tập Kvasir và đứng thứ 2 trên tập CVC-ClinicDB, tuy nhiên chỉ thấp hơn không đáng kể so với mô hình tốt nhất là TransFuse-L* 0.5% ở độ đo mDice. Điều này cho thấy mô hình đề xuất có khả năng học để phân vùng hiệu quả polyp trên ảnh nội soi.

Bảng 4.4: So sánh hiệu năng của các phương pháp trên 5 tập kiểm tra: Kvasir, ClinicDB, ColonDB, EndoScene (CVC-T) và ETIS

Phương pháp	Kvasir		ClinicDB		ColonDB		CVC-T		ETIS	
	mDice	mIoU								
UNet [15]	0.818	0.746	0.823	0.755	0.512	0.444	0.710	0.627	0.398	0.335
UNet++ [16]	0.821	0.743	0.794	0.729	0.483	0.410	0.707	0.624	0.401	0.344
SFA [29]	0.723	0.611	0.700	0.607	0.469	0.347	0.467	0.329	0.297	0.217
PraNet [2]	0.898	0.840	0.899	0.849	0.709	0.640	0.871	0.797	0.628	0.567
HarDNet-MSEG [19]	0.912	0.857	0.932	0.882	0.731	0.660	0.887	0.821	0.677	0.613
CaraNet [20]	0.918	0.865	0.936	0.887	0.773	0.689	0.903	0.838	0.747	0.672
TransUnet [3]	0.913	0.857	0.935	0.887	0.781	0.699	0.893	0.824	0.731	0.660
TransFuse-L* [4]	0.920	0.870	0.942	0.897	0.781	0.706	0.894	0.826	0.737	0.663
Polyp-PVT [5]	0.917	0.864	0.937	0.889	0.808	0.727	0.900	0.833	0.787	0.706
Mô hình đề xuất	0.926	0.878	0.937	0.890	0.813	0.734	0.895	0.827	0.803	0.725
±	0.004	0.004	0.003	0.003	0.011	0.010	0.004	0.006	0.006	0.007

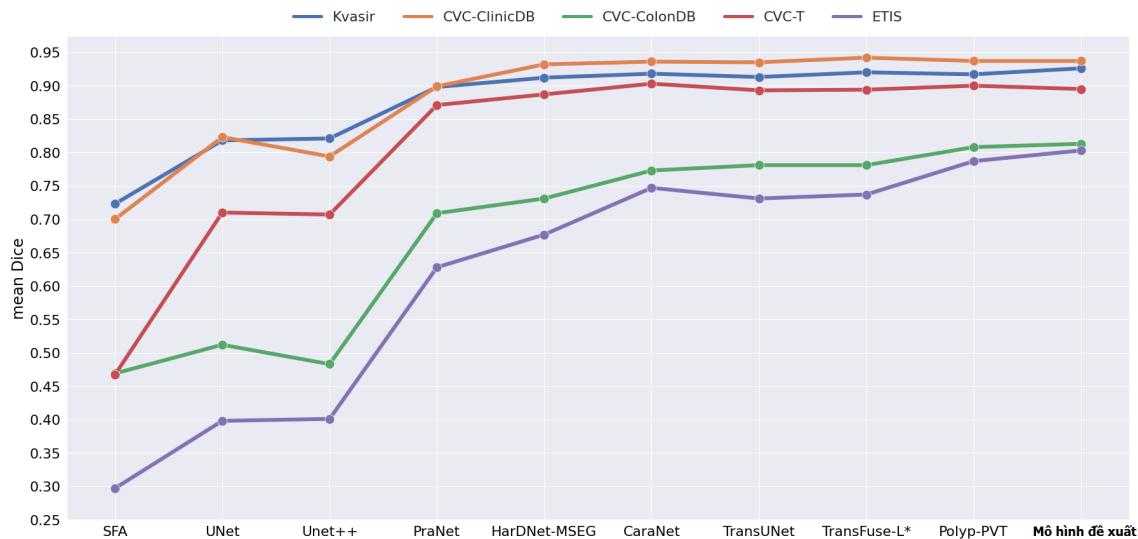
Kết quả của các phương pháp ngoài mô hình đề xuất có tham khảo tại CaraNet [20] và TransFuse [4].

Kết quả của mô hình đề xuất bao gồm giá trị trung bình và độ lệch chuẩn của 5 lần chạy.

Từ Bảng 4.4 cũng có thể nhận thấy rằng, độ chính xác của mô hình đề xuất cao hơn các mô hình khác trên hai tập CVC-ColonDB, ETIS và xếp thứ 3 trên tập CVC-T. Ngoài ra, trên tập CVC-ColonDB và ETIS, mô hình cho kết quả đứng thứ hai là Polyp-PVT, đây cũng là một mô hình có kiến trúc dựa trên Transformer. Cũng trên hai tập này, CaraNet là mô hình cho kết quả cao nhất trong nhóm các mô hình thuần CNN. Cụ thể, với tập CVC-ColonDB, mô hình đề xuất cho kết quả cao hơn 0.5% trên độ đo mDice và 0.7% trên độ đo mIoU so với Polyp-PVT; so với CaraNet, mô hình đề xuất cho kết quả cao hơn 4% trên độ đo mDice và 4.5% trên độ đo mIoU. Với tập ETIS, mô hình đề xuất cho kết quả vượt trội, cao hơn 1.6% trên mDice và 1.9% trên mIoU so với Polyp-PVT và cao hơn 5.6% trên mDice và 5.3% trên mIoU so với CaraNet. Với tập CVC-T, mô hình đề xuất có độ chính xác thấp hơn 0.8% trên mDice và 1.1% trên mIoU so với mô hình tốt nhất là CaraNet.

Kết quả thực nghiệm này đã cho thấy mô hình đề xuất có khả năng tổng quát hóa tốt trên các bộ dữ liệu khác nhau, đồng thời cũng cho thấy khả năng trích xuất và biểu diễn đặc trưng mạnh mẽ của các kiến trúc dựa trên Transformer.

Khả năng học và tổng quát hóa của các phương pháp cũng được trình bày trực quan ở Hình 4.4 trên độ đo mean Dice. Có thể thấy, phương pháp đề xuất đã đạt được những cải thiện đáng kể so với các phương pháp hiện tại.



Hình 4.4: So sánh độ chính xác các phương pháp khác nhau.

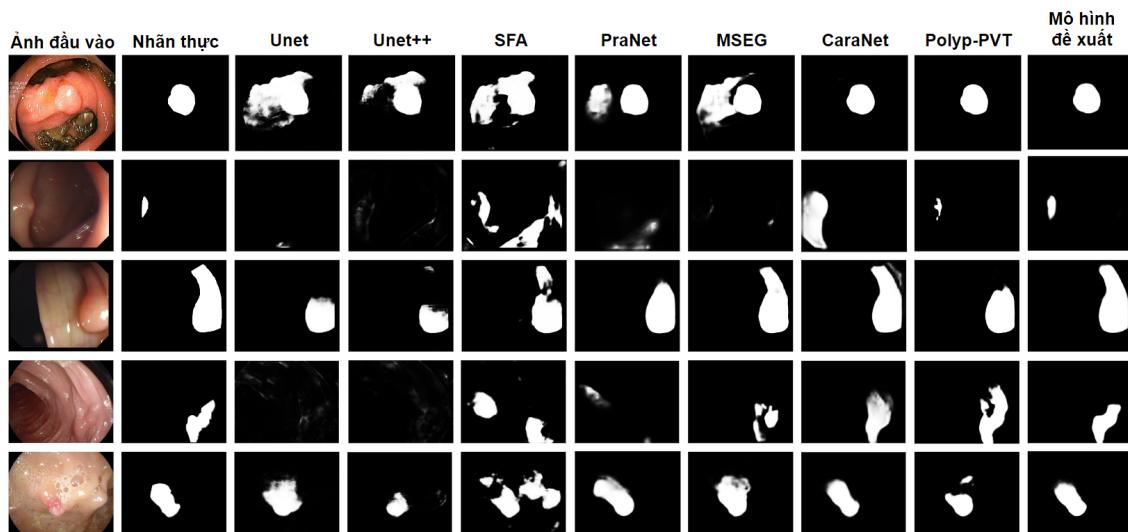
Bảng 4.5 so sánh số lượng tham số và độ phức tạp tính toán của mô hình đề xuất với các mô hình khác. Có thể nhận thấy rằng, mô hình đề xuất có số lượng tham số và độ phức tạp đều nhỏ hơn các mô hình dựa trên Transformer. Với các mô hình thuần CNN, so với CaraNet, mô hình đề xuất có số lượng tham số và độ phức tạp đều nhỏ hơn; so với PraNet và HarDNet-MSEG, tuy hai thông số này của mô hình đề xuất đều lớn hơn nhưng khả năng học và khả năng tổng quát hóa của mô hình đề xuất đều vượt trội hơn hai mô hình này.

Bảng 4.5: Số lượng tham số và độ phức tạp của các phương pháp khác nhau

Phương pháp	Số lượng tham số (M)	GFLOPs
PraNet [2]	32.55	13.18
HarDNet-MSEG [19]	33.34	11.41
CaraNet [20]	46.64	21.80
TransUNet [3]	105.5	60.75
TransFuse-L* [4]	—	—
Polyp-PVT [5]	—	—
MiT-PD (MiT + PD decoder)	45.62	21.61
MiT-MLP (MiT + Segformer decoder)	45.25	20.19
Mô hình đề xuất	45.68	20.72

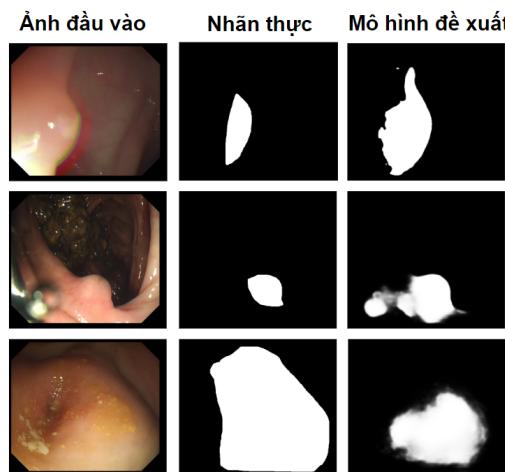
Các thông số của TransUNet có tham khảo tại ColonFormer [30].

Hình 4.5 hiển thị trực quan kết quả dự đoán của mô hình đề xuất và các mô hình khác trên một số ảnh nội soi polyp phức tạp, chẳng hạn ảnh có độ tương phản thấp giữa vùng chứa polyp và vùng niêm mạc xung quanh, ảnh bị nhòe, ảnh có vùng đổ bóng, hay có kết cấu đa dạng, ... Có thể nhận thấy rằng, mô hình đề xuất có hai ưu điểm so với các mô hình khác: (i) Mô hình có khả năng nhận diện và phân vùng ổn định, thích ứng với dữ liệu trong các điều kiện khác nhau, chẳng hạn trong các điều kiện ánh sáng đa dạng, độ tương phản, phản chiếu khác nhau. (ii) Mô hình đề xuất cho ra kết quả phân vùng ít nhiễu hơn so với các phương pháp khác. Đồng thời, các kết quả phân vùng của mô hình có tính nhất quán bên trong và vùng biên của polyp được dự đoán gần với nhãn thực tế hơn.



Hình 4.5: Một số kết quả dự đoán của mô hình đề xuất và các mô hình khác.

Mặc dù mô hình đề xuất đạt được một số cải thiện đáng kể so với các mô hình trước đây, tuy nhiên vẫn tồn tại một số trường hợp cho hiệu năng không được tốt. Hình 4.6 hiển thị kết quả dự đoán của một số trường hợp hạn chế này.



Hình 4.6: Một số nhược điểm của mô hình đề xuất.

Có thể thấy, một trong những hạn chế tồn tại của mô hình đề xuất là không phát hiện chính xác ranh giới polyp tại nơi giao thoa giữa vùng có ánh sáng chiếu đèn và vùng có bóng tối. Như trong Hình 4.6 ở hàng thứ nhất, ngoài phần vùng được polyp thực, mô hình còn nhận nhầm nơi giao thoa giữa vùng sáng và tối như là polyp. Ở hàng 2, ảnh chứa nhiễu, nhiễu này có thể được tạo ra do hiện tượng nhòe chuyển động hoặc hiện tượng phản xạ ánh sáng, loại nhiễu này làm mô hình nhận nhầm là polyp. Đây là một trong những nguyên nhân ảnh hưởng đến độ chính xác trong tập CVC-ColonDB và CVC-T. Ở hàng 3, diện tích polyp chiếm phần lớn trong ảnh nhưng kết quả dự đoán phân vùng polyp lại có diện tích nhỏ hơn nhiều. Nguyên nhân được suy đoán cho trường hợp này là bởi mô hình đề xuất sử dụng bộ khung trích xuất đặc trưng dựa trên Transformer, trích xuất ra các đặc trưng có tính toàn cục. Tuy nhiên với những trường hợp mà diện tích polyp chiếm phần lớn trong ảnh, mô hình có thể không trích xuất được đủ thông tin ngữ cảnh cần thiết cho quá trình dự đoán.

4.5 Đánh giá ảnh hưởng của các mô-đun

Các mô hình được trình bày sau đây sử dụng chung chiến lược huấn luyện như được đề cập trong Mục 4.3. Việc huấn luyện các mô hình được thực hiện 5 lần, kết quả thí nghiệm thu được bằng cách tính trung bình các kết quả qua 5 lần chạy.

Đánh giá hiệu quả của mô-đun giải mã MLP: em so sánh mô hình MiT-MLP với mô hình khác được ký hiệu MiT-PD. Trong đó cả hai mô hình đều có phần mã hóa là kiến trúc MiT, tuy nhiên MiT-MLP có phần giải mã là kiến trúc MLP được đề xuất trong bài báo SegFormer [7] và MiT-PD có phần giải mã PD được đề xuất trong PraNet [2]. Kết quả được trình bày trong Bảng 4.6 cho thấy rằng cả hai mô hình đều cho hiệu năng tương đồng nhau trên khả năng học và khả năng tổng quát hóa, với sự sai khác độ chính xác không quá 1%. Tuy nhiên, như được thống kê trong Bảng 4.5, số lượng tham số và độ phức tạp tính toán của mô-đun PD đều lớn hơn MLP. Do đó, để cân bằng giữa độ chính xác và độ phức tạp tính toán, trong đồ án sử dụng kiến trúc MLP cho mô-đun giải mã.

Đánh giá hiệu quả của mô-đun CBAM: để đánh giá hiệu quả của mô-đun CBAM, em so sánh mô hình MiT-MLP và mô hình khác có tên MiT-CBAM-MLP, mô hình này có tích hợp thêm CBAM trong mô-đun giải mã MLP như được trình bày trong Mục 3.3.1. Kết quả được trình bày trong Bảng 4.6. Có thể thấy, với sự hỗ trợ của CBAM, hiệu năng của mô hình được cải thiện trên cả khả năng học và khả năng tổng quát hóa.

Đánh giá hiệu quả của mô-đun tinh chỉnh đặc trưng: em so sánh mô hình MiT-CBAM-MLP với mô hình khác có tên MiT-CBAM-MLP-RA, mô hình này có

tích hợp thêm mô-đun tinh chỉnh đặc trưng như được trình bày trong Mục 3.4. Kết quả được trình bày trong Bảng 4.6, có thể thấy mô-đun tinh chỉnh tại cả 4 đặc trưng cho ra bởi MiT giúp tăng hiệu năng của mô hình trên tất cả các bộ dữ liệu kiểm tra, đặc biệt tăng 1.3% ở cả độ đo mDice và mIoU trên tập ETIS. Hình 4.8 hiển thị kết quả dự đoán tại 5 đầu ra biên của mô hình, bao gồm bản đồ đặc trưng toàn cục S_g và S_4, S_3, S_2, S_1 là 4 đầu ra được tinh chỉnh tại 4 đặc trưng cho ra bởi MiT. Trong đó S_4 là đầu ra được tinh chỉnh tại đặc trưng mức cao nhất, S_1 là đầu ra được tinh chỉnh tại đặc trưng mức thấp nhất và cũng là kết quả dự đoán cuối cùng của mô hình. Có thể nhận thấy rằng, bản đồ đặc trưng toàn cục S_g đã xác định được polyp tương đối hoàn chỉnh, nhưng thường chứa nhiều nhiễu. Với sự hỗ trợ của mô-đun tinh chỉnh đặc trưng, các đầu ra dự đoán dần dần được cải thiện, loại bỏ bớt nhiễu và tinh chỉnh kết quả phân vùng tại biên của polyp.

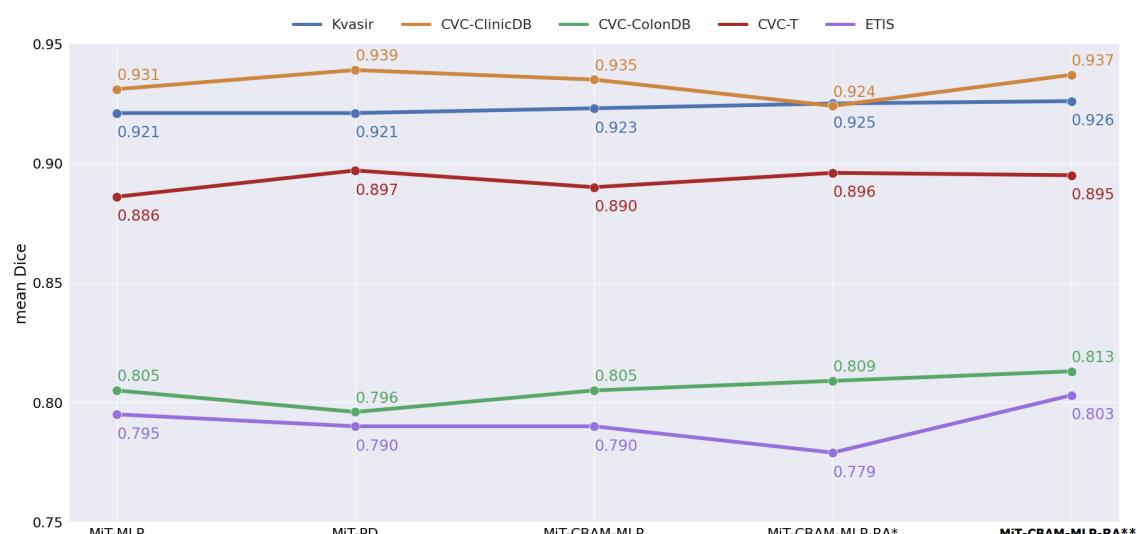
Bảng 4.6: Ảnh hưởng của các mô-đun đến khả năng học và khả năng tổng quát hóa của mô hình

Phương pháp	CBAM	RA	MLP	PD	Kvasir mDice mIoU	ClinicDB mDice mIoU	ColonDB mDice mIoU	CVC-T mDice mIoU	ETIS mDice mIoU
MiT-MLP	—	—	✓	—	0.921 0.872	0.931 0.882	0.805 0.726	0.886 0.819	0.795 0.716
MiT-PD	—	—	—	✓	0.921 0.872	0.939 0.893	0.796 0.717	0.897 0.828	0.790 0.708
MiT-CBAM-MLP	✓	—	✓	—	0.923 0.874	0.935 0.887	0.805 0.725	0.890 0.822	0.790 0.712
MiT-CBAM-MLP-RA*	✓	✓	✓	—	0.925 0.877	0.924 0.875	0.809 0.731	0.896 0.828	0.779 0.700
MiT-CBAM-MLP-RA**	✓	✓	✓	—	0.926 0.878	0.937 0.890	0.813 0.734	0.895 0.827	0.803 0.725

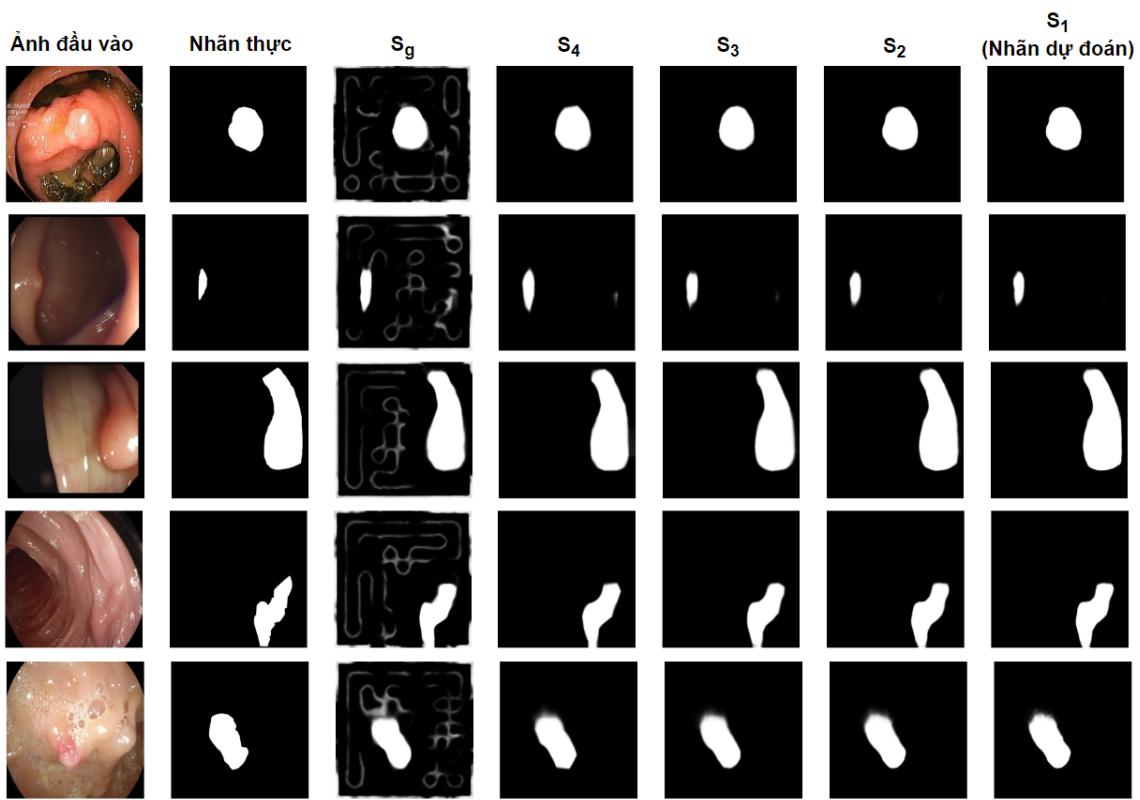
*: tinh chỉnh tại 3 đặc trưng cho ra bởi MiT

**: tinh chỉnh tại cả 4 đặc trưng cho ra bởi MiT

Ảnh hưởng của các mô-đun cũng được trình bày trực quan trên Hình 4.7 ở độ đo mean Dice. Có thể thấy, mô hình đã xuất đã kết hợp hiệu quả các mô-đun để nâng cao hiệu năng tổng thể trên cả khả năng học và khả năng tổng quát hóa.



Hình 4.7: Ảnh hưởng của các mô-đun đến độ chính xác của mô hình.



Hình 4.8: Kết quả dự đoán tại các đầu ra biên của mô hình đề xuất.

4.6 Đánh giá ảnh hưởng của các phương pháp khởi tạo trọng số

Trong quá trình huấn luyện mô hình đề xuất, ngoại trừ trọng số khởi tạo của mô-đun mã hóa MiT sử dụng trọng số được tiền huấn luyện trên bộ dữ liệu ImageNet, hai mô-đun còn lại là mô-đun giải mã và mô-đun tinh chỉnh đặc trưng sử dụng phương pháp khởi tạo He normal. Bảng 4.7 trình bày sự ảnh hưởng của các phương pháp khởi tạo trọng số thông qua độ đo mean Dice và giá trị độ lệch chuẩn của 5 lần chạy. Có thể thấy rằng, mô hình đề xuất sử dụng phương pháp khởi tạo He normal cho kết quả cao và ổn định hơn các phương pháp khởi tạo khác.

Bảng 4.7: Ảnh hưởng của các phương pháp khởi tạo trọng số

Phương pháp khởi tạo	Kvasir mDice ± SD	ClinicDB mDice ± SD	ColonDB mDice ± SD	CVC-T mDice ± SD	ETIS mDice ± SD
Ngẫu nhiên	.919 ± .006	.933 ± .012	.808 ± .009	.889 ± .009	.805 ± .014
Xavier normal	.919 ± .006	.921 ± .006	.820 ± .011	.892 ± .012	.806 ± .016
He normal	.926 ± .004	.937 ± .003	.813 ± .011	.895 ± .004	.803 ± .006

Kết chương: trong chương 4 đã trình bày kết quả thực nghiệm, so sánh, đánh giá định lượng và định tính mô hình đề xuất trên khả năng học và khả năng tổng quát hóa với các mô hình khác. Đồng thời, cho thấy sự hiệu quả của các mô-đun được tích hợp và phương pháp khởi tạo trọng số được sử dụng khi huấn luyện mô

hình đề xuất. Chương sau sẽ rút ra kết luận, nêu lên ưu, nhược điểm của mô hình và đề ra hướng phát triển trong tương lai.

CHƯƠNG 5. KẾT LUẬN

5.1 Kết luận

Trên đây em đã trình bày nội dung đồ án “Phân vùng Polyp trên ảnh nội soi dựa trên kiến trúc Transformer” mà em đã thực hiện dưới sự hướng dẫn của TS. Nguyễn Thị Oanh. Đồ án đã hoàn thành các yêu cầu đề ra và đạt được một số kết quả nhất định. Vì thời gian nghiên cứu, tìm hiểu ngắn và còn nhiều hạn chế về trang thiết bị huấn luyện mô hình nên trong đồ án chỉ hoàn thành ở mức cơ bản và không thể tránh khỏi một vài thiếu sót khi trình bày và đánh giá vấn đề. Rất mong nhận được sự góp ý, đánh giá của các thầy cô bộ môn để đề tài của em thêm hoàn thiện hơn.

Trong đồ án đã thực hiện tìm hiểu kiến trúc và xây dựng mô hình mạng nơron có khả năng phân vùng polyp từ ảnh nội soi đại tràng đạt độ chính xác cao, kích thước lưu trữ nhỏ và độ phức tạp tính toán thấp. Mô hình được xây dựng là sự kết hợp của ba thành phần: mô-đun mã hóa trích xuất đặc trưng ảnh đầu vào sử dụng Mix Transformer, mô-đun giải mã tạo bản đồ đặc trưng toàn cục bằng cách kết hợp song song các đặc trưng với sự hỗ trợ của cơ chế chú ý theo cả chiều sâu và chiều không gian (CBAM), mô-đun tinh chỉnh đặc trưng là sự kết hợp giữa trích xuất đặc trưng kim tự tháp theo kênh (CFP) và cơ chế chú ý ngược (RA).

Kết quả thực nghiệm cho thấy mô hình đề xuất đạt kết quả tốt trên cả 5 tập dữ liệu đánh giá mà không cần sử dụng bất kỳ phép hậu xử lý nào. Đồng thời, đạt độ chính xác cao không những trên các mẫu dữ liệu ảnh polyp có kích thước lớn mà còn trên các các mẫu dữ liệu chứa polyp có kích thước nhỏ hoặc rất nhỏ.Thêm vào đó, mô hình mạng đề xuất có số lượng tham số và số lượng phép tính dấu phẩy động (FLOPs) đều thấp hơn so với các kiến trúc mạng SOTA gần đây cho bài toán phân vùng polyp.

Mặc dù đạt được những cải tiến đáng kể nhưng mô hình vẫn còn tồn tại một số hạn chế có thể kể đến như: chưa xử lý được hiện tượng nhận nhầm là polyp trong các trường hợp ảnh nội soi có nhiễu, xảy ra vấn đề nhòe chuyển động hay vùng niêm mạc bình thường nhưng có ánh sáng phản chiếu đèn. Độ chính xác phân vùng chưa cao ở những vùng đổ bóng hay vùng giao giữa ánh sáng và bóng tối.

5.2 Hướng phát triển trong tương lai

Mô hình đề xuất tồn tại một số tiềm năng đáng kể để tối ưu hóa mô hình. Chẳng hạn, sử dụng toán tử nội suy song tuyến để thay đổi kích thước đặc trưng dẫn đến không thể tránh khỏi xảy ra mất mát thông tin. Có thể cải tiến bằng cách thay thế toán tử nội suy song tuyến bằng một tầng tích chập có kích thước bộ lọc phù hợp.

Ngoài ra, bộ khung trích xuất đặc trưng MiT sử dụng trọng số khởi tạo được tiền huấn luyện trên bộ dữ liệu ImageNet, đây là bộ dữ liệu chứa ảnh tự nhiên nên có tính chất khá khác với ảnh nội soi y tế.

Trong tương lai để cải tiến hơn nữa, có thể thiết kế chiến lược tăng cường dữ liệu, nghiên cứu cách kết hợp mô hình đề xuất với các phương pháp học bán giám sát để giải quyết vấn đề khan hiếm nguồn dữ liệu y tế hay các phương pháp học thích ứng miên để nâng cao khả năng tổng quát hóa của mô hình trên các miền dữ liệu đa dạng, chẳng hạn trên các hệ thống máy, các chế độ ánh sáng nội soi khác nhau.

TÀI LIỆU THAM KHẢO

- [1] D. A. Corley, C. D. Jensen, A. R. Marks, *et al.*, “Adenoma detection rate and risk of colorectal cancer and death,” *New england journal of medicine*, vol. 370, no. 14, pp. 1298–1306, 2014.
- [2] D.-P. Fan, G.-P. Ji, T. Zhou, *et al.*, “Pranet: Parallel reverse attention network for polyp segmentation,” in *International conference on medical image computing and computer-assisted intervention*, Springer, 2020, pp. 263–273.
- [3] J. Chen, Y. Lu, Q. Yu, *et al.*, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.
- [4] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2021, pp. 14–24.
- [5] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, “Polyp-pvt: Polyp segmentation with pyramid vision transformers,” *preprint arXiv:2108.06932*, 2021.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *preprint arXiv:2010.11929*, 2020.
- [7] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, “Segformer: Simple and efficient design for semantic segmentation with transformers,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.
- [8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*, Springer, 2020, pp. 213–229.
- [9] W. Wang, E. Xie, X. Li, *et al.*, “Pvt v2: Improved baselines with pyramid vision transformer,” *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.
- [10] Z. Liu, Y. Lin, Y. Cao, *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
- [11] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

- [12] M. Fiori, P. Musé, and G. Sapiro, “A complete system for candidate polyps detection in virtual colonoscopy,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 28, no. 07, p. 1460014, 2014.
- [13] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, “Automated polyp detection in colon capsule endoscopy,” *IEEE transactions on medical imaging*, vol. 33, no. 7, pp. 1488–1502, 2014.
- [14] O. H. Maghsoudi, “Superpixel based segmentation and classification of polyps in wireless capsule endoscopy,” in *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, IEEE, 2017, pp. 1–4.
- [15] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.
- [16] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, Springer, 2018, pp. 3–11.
- [17] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, “Doubleunet: A deep convolutional neural network for medical image segmentation,” in *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, IEEE, 2020, pp. 558–564.
- [18] D. Jha, P. H. Smedsrød, M. A. Riegler, *et al.*, “Resunet++: An advanced architecture for medical image segmentation,” in *2019 IEEE International Symposium on Multimedia (ISM)*, IEEE, 2019, pp. 225–2255.
- [19] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, “Hardnet-mseg: A simple encoder-decoder polyp segmentation neural network that achieves over 0.9 mean dice and 86 fps,” *preprint arXiv:2101.07172*, 2021.
- [20] A. Lou, S. Guan, and M. Loew, “Caranet: Context axial reverse attention network for segmentation of small medical objects,” *preprint arXiv:2108.07368*, 2021.
- [21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
- [22] A. Lou and M. Loew, “Cfpnet: Channel-wise feature pyramid for real-time semantic segmentation,” in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 1894–1898.

- [23] S. Chen, X. Tan, B. Wang, and X. Hu, “Reverse attention for salient object detection,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 234–250.
- [24] D. Jha, P. H. Smedsrud, M. A. Riegler, *et al.*, “Kvasir-seg: A segmented polyp dataset,” in *International Conference on Multimedia Modeling*, Springer, 2020, pp. 451–462.
- [25] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.
- [26] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated polyp detection in colonoscopy videos using shape and context information,” *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 630–644, 2015.
- [27] D. Vázquez, J. Bernal, F. J. Sánchez, *et al.*, “A benchmark for endoluminal scene segmentation of colonoscopy images,” *Journal of healthcare engineering*, vol. 2017, 2017.
- [28] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer,” *International journal of computer assisted radiology and surgery*, vol. 9, no. 2, pp. 283–293, 2014.
- [29] Y. Fang, C. Chen, Y. Yuan, and K.-y. Tong, “Selective feature aggregation network with area-boundary constraints for polyp segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2019, pp. 302–310.
- [30] N. T. Duc, N. T. Oanh, N. T. Thuy, T. M. Triet, and D. V. Sang, “Colonformer: An efficient transformer based method for colon polyp segmentation,” *arXiv preprint arXiv:2205.08473*, 2022.