

**ĐẠI HỌC BÁCH KHOA HÀ NỘI**

**ĐỒ ÁN TỐT NGHIỆP**

**Mô hình nền tảng cho ảnh nội soi y tế**

**LÊ ĐÌNH TUYÊN**

tuyen.ld194715@sis.hust.edu.vn

**Ngành Công nghệ thông tin**

**Giảng viên hướng dẫn:** TS. Đinh Viết Sang

Chữ ký GVHD

**Khoa:** Khoa học máy tính

**Trường:** Công nghệ Thông tin và Truyền thông

**HÀ NỘI, 06/2024**

# LỜI CẢM ƠN

Trong quá trình thực hiện và hoàn thành đồ án này, em xin bày tỏ lòng biết ơn sâu sắc đến những người đã hỗ trợ và đồng hành cùng em.

Đầu tiên, em xin gửi lời cảm ơn đến thầy Đinh Việt Sang. Nhờ sự hướng dẫn tận tình, sự chỉ bảo chu đáo, luôn quan tâm đến sinh viên của thầy, em đã có thể tiếp thu và áp dụng những kiến thức quý báu đó vào thực hiện đồ án này. Sự quan tâm và hỗ trợ của thầy là nguồn động lực lớn giúp em hoàn thành đồ án tốt nghiệp.

Tiếp theo, con cũng xin gửi lời cảm ơn đến bố mẹ, gia đình và những người thân yêu đã luôn động viên và ủng hộ trong suốt quá trình học tập và thực hiện đồ án. Quãng thời gian học đại học tuy có vất vả, khó khăn nhưng bằng chính những lời động viên đó đã giúp con vượt qua tất cả để đi đến cuối cùng của đoạn đường này.

Bên cạnh đó, tôi cũng không quên gửi lời cảm ơn đến bạn bè, những người đã hỗ trợ, chia sẻ kiến thức giúp tôi vượt qua những thử thách trong suốt quãng đời sinh viên với bao niềm vui, nỗi buồn. Việc được gặp gỡ các bạn đã làm cho quãng đời sinh viên của tôi trở nên rực rỡ và khó quên hơn bao giờ hết.

Cuối cùng, mặc dù đã cố gắng hết sức, nhưng do vốn kiến thức và kinh nghiệm còn hạn chế, đồ án này khó tránh khỏi những thiếu sót. Em mong nhận được những góp ý và phản hồi từ thầy cô để em có thể hoàn thiện hơn trong tương lai.

Em xin chân thành cảm ơn!

# LỜI CAM KẾT

Họ và tên sinh viên: Lê Đình Tuyên  
Điện thoại liên lạc: 0824383925  
Email: le.tuyen.hust@gmail.com  
Lớp: IS-03 K64  
Hệ đào tạo: Công nghệ thông tin Việt Nhật

Tôi – *Lê Đình Tuyên* – cam kết Đồ án Tốt nghiệp (ĐATN) là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *TS. Đinh Viết Sang*. Các kết quả nêu trong ĐATN là trung thực, là thành quả của riêng tôi, không sao chép theo bất kỳ công trình nào khác. Tất cả những tham khảo trong ĐATN – bao gồm hình ảnh, bảng biểu, số liệu, và các câu trích dẫn – đều được ghi rõ ràng và đầy đủ nguồn gốc trong danh mục tài liệu tham khảo. Tôi xin hoàn toàn chịu trách nhiệm với dù chỉ một sao chép vi phạm quy chế của nhà trường.

*Hà Nội, ngày 30 tháng 06 năm 2024*

Tác giả ĐATN

*Lê Đình Tuyên*

# TÓM TẮT NỘI DUNG ĐỒ ÁN

Trong kỷ nguyên số hóa hiện nay, lượng dữ liệu mà chúng ta tạo ra và thu thập được đang ngày càng tăng một cách chóng mặt. Tuy nhiên, trong dữ liệu khổng lồ này, một vấn đề lớn đang nổi lên: phần lớn dữ liệu không được gán nhãn. Dữ liệu thiếu nhãn là một trong những thách thức lớn nhất trong trí tuệ nhân tạo. Trong nhiều trường hợp, việc thu thập dữ liệu có nhãn là tốn kém và tốn thời gian. Đặc biệt là trong lĩnh vực y tế, khi cần có những bác sĩ có chuyên môn cao gán nhãn cho những hình ảnh y tế. Điều đó gây ra sự lãng phí công sức, thời gian, tiền bạc. Để giải quyết vấn đề này, các phương pháp tự giám sát đã cho thấy tiềm năng to lớn nhờ sử dụng một lượng lớn dữ liệu không có nhãn. Trong các phương pháp tự giám sát, phương pháp sử dụng chiến lược che đi một phần của ảnh như BeIT [1], iBot [2], SimMIM [3], ... đã chứng minh được hiệu quả vượt trội bằng cách yêu cầu mô hình hiểu và nắm bắt được cấu trúc cũng như các đặc điểm quan trọng của hình ảnh. Trong đồ án này, chúng tôi đề xuất sử dụng mô hình Masked Autoencoder (MAE) [4], một mô hình che hình ảnh để huấn luyện tự giám sát với dữ liệu ảnh nội soi y tế không nhãn nhằm tạo một xương sống mạnh mẽ cho các nhiệm vụ mục tiêu như phân vùng tổn thương, phân loại tổn thương trên hình ảnh nội soi y tế. Kết quả thử nghiệm trên các tập dữ liệu khác nhau cho thấy, trong giai đoạn tinh chỉnh với cùng một lượng dữ liệu có nhãn, mô hình được tiếp tục tiền huấn luyện trên dữ liệu ảnh nội soi y tế cho hiệu suất tốt hơn mô hình được tiền huấn luyện từ đầu trên dữ liệu ảnh nội soi y tế hoặc mô hình chỉ được tiền huấn luyện với ImageNet1K [5]. Bên cạnh đó, mô hình của chúng tôi cũng đạt hiệu suất gần ngang bằng so với các mô hình tiên tiến nhất (state of the art) về phân vùng polyp. Ngoài việc đề xuất mô hình đơn nhiệm vụ, chúng tôi cũng đề xuất một mô hình sử dụng bộ mã hóa của Masked Autoencoders làm xương sống cho mô hình thực hiện nhiều nhiệm vụ và qua các kết quả thử nghiệm, mô hình của chúng tôi cho hiệu suất cao hơn so với mô hình cơ sở về thực hiện đa nhiệm vụ.

Sinh viên thực hiện

(Lê Đình Tuyên)

## MỤC LỤC

<b>CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI.....</b>	<b>1</b>
1.1 Đặt vấn đề.....	1
1.2 Phạm vi của đồ án.....	2
1.3 Đóng góp của đồ án .....	3
1.4 Bố cục đồ án .....	3
<b>CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT .....</b>	<b>4</b>
2.1 Khái niệm về học máy và học sâu.....	4
2.2 Các phương pháp học máy.....	4
2.3 Các mô hình học sâu .....	7
2.3.1 Mạng nơ-ron nhân tạo .....	7
2.3.2 Mạng nơ-ron tích chập.....	7
2.3.3 Transformers .....	8
2.3.4 Vision Transformers .....	11
2.4 Bài toán phân vùng ảnh y tế và các mô hình học sâu .....	12
2.4.1 Bài toán phân vùng ảnh .....	12
2.4.2 Giám sát sâu (Deep supervision) .....	12
2.4.3 U-Net.....	12
2.4.4 PraNet.....	13
2.5 Mô hình đa nhiệm.....	14
<b>CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT.....</b>	<b>17</b>
3.1 Tổng quan .....	17
3.2 Tiền huấn luyện.....	17
3.3 Tinh chỉnh trên các nhiệm vụ mục tiêu .....	22
3.3.1 Mô hình phân vùng hình ảnh nội soi y tế .....	22

3.3.2 Mô hình đa nhiệm .....	25
<b>CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM.....</b>	<b>27</b>
4.1 Dữ liệu .....	27
4.2 Tiền xử lý dữ liệu và tăng cường dữ liệu.....	29
4.3 Triển khai chi tiết.....	31
4.4 Các độ đo.....	31
4.4.1 Độ đo Dice .....	32
4.4.2 Độ đo IoU .....	33
4.4.3 Độ chính xác .....	34
4.5 Kết quả thử nghiệm.....	34
4.5.1 Chất lượng tái tạo hình ảnh .....	34
4.5.2 Phân đoạn hình ảnh nội soi .....	35
4.5.3 Đánh giá trên các điểm chuẩn Polyp .....	37
4.5.4 Mô hình đa nhiệm .....	38
<b>CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN .....</b>	<b>43</b>
5.1 Kết luận .....	43
5.2 Hướng phát triển trong tương lai .....	43
<b>TÀI LIỆU THAM KHẢO.....</b>	<b>47</b>

## DANH MỤC HÌNH VẼ

Hình 2.1	Huấn luyện mô hình tự giám sát với nhiệm vụ giả định là các phép xoay ảnh. <sup>1</sup>	6
Hình 2.2	Mạng học sâu nơ-ron được lấy cảm hứng từ não người. <sup>2</sup>	7
Hình 2.3	Kiến trúc mạng nơ-ron tích chập - CNN [12]	8
Hình 2.4	Kiến trúc Transformers [8]	9
Hình 2.5	Kiến trúc Vision Transformer [9]	11
Hình 2.6	Phân vùng ngữ nghĩa và phân vùng đối tượng <sup>3</sup>	12
Hình 2.7	Kiến trúc UNet [16]	13
Hình 2.8	Kiến trúc của mô hình PraNet [17]	14
Hình 2.9	Minh họa mô hình đa nhiệm <sup>4</sup>	14
Hình 2.10	Kiến trúc EndoUNet [19]	15
Hình 3.1	Thiết kế tổng quan sử dụng MAE [4] cho quá trình tiền huấn luyện và tinh chỉnh	17
Hình 3.2	Kiến trúc mô hình Masked Autoencoders [4]	18
Hình 3.3	Hiển thị trực quan một số mảng nền gây ra sự chú ý trong DEiT [21], CLIP [22], DINOv2 [23] mà các nhà nghiên cứu muốn giảm bớt.	19
Hình 3.4	Ảnh hưởng của các token bổ sung (register token) giúp ViT chú ý chính xác đến đối tượng	21
Hình 3.5	Minh họa cho việc thêm register token vào MAE [4]	21
Hình 3.6	Kiến trúc mô hình phân vùng hình ảnh nội soi y tế	23
Hình 3.7	Kiến trúc mô hình đa nhiệm	25
Hình 4.1	Một số kỹ thuật tăng cường dữ liệu được sử dụng	30
Hình 4.2	Lập lịch khởi động tốc độ học Cosine	31
Hình 4.3	Minh họa chất lượng tái tạo hình ảnh của MAE [4]	35
Hình 4.4	Một số hình ảnh dự đoán của các mô hình được huấn luyện chỉ với 5% dữ liệu và suy luận trên tập thử nghiệm	40
Hình 4.5	Ma trận nhầm lẫn của phân loại HP sử dụng xương sống của MAE	41
Hình 4.6	Ma trận nhầm lẫn của phân loại vị trí giải phẫu trên mô hình đa nhiệm sử dụng xương sống của MAE	41
Hình 4.7	Ma trận nhầm lẫn của phân loại bệnh trên mô hình đa nhiệm sử dụng xương sống của MAE	42

## **DANH MỤC BẢNG BIỂU**

Bảng 4.1	Dữ liệu được trích xuất từ video . . . . .	27
Bảng 4.2	Dữ liệu huấn luyện mô hình PraNet [17] . . . . .	27
Bảng 4.3	Dữ liệu kiểm thử mô hình PraNet [17] . . . . .	28
Bảng 4.4	Dữ liệu polyp từ bệnh viện . . . . .	28
Bảng 4.5	Dữ liệu tiền huấn luyện . . . . .	28
Bảng 4.6	Dữ liệu huấn luyện và kiểm thử phân loại vị trí giải phẫu . . .	29
Bảng 4.7	Dữ liệu tiền huấn luyện . . . . .	29
Bảng 4.8	Cài đặt siêu tham số cho tiếp tục tiền huấn luyện. . . . .	32
Bảng 4.9	Cài đặt siêu tham số cho các nhiệm vụ mục tiêu. . . . .	32
Bảng 4.10	Kết quả thử nghiệm của các mô hình khi huấn luyện trên từng tập dữ liệu với 5% dữ liệu và kiểm thử trên tập kiểm thử tương ứng .	36
Bảng 4.11	Kết quả thử nghiệm của các mô hình khi huấn luyện trên từng tập dữ liệu với 10% dữ liệu và kiểm thử trên tập kiểm thử tương ứng. .	36
Bảng 4.12	Kết quả thử nghiệm của các mô hình khi huấn luyện trên từng tập dữ liệu và kiểm thử trên tập kiểm thử tương ứng. . . . .	36
Bảng 4.13	Kết quả thử nghiệm của các mô hình trên các điểm chuẩn polyp. . . . .	37
Bảng 4.14	Kết quả thử nghiệm phân loại trên các mô hình đa nhiệm. . . .	38
Bảng 4.15	Kết quả thử nghiệm của các mô hình đa nhiệm trên nhiệm vụ phân vùng . . . . .	38

## **DANH MỤC THUẬT NGỮ VÀ TỪ VIẾT TẮT**

<b>Thuật ngữ</b>	<b>Ý nghĩa</b>
ANN	Mạng nơ-ron nhân tạo
CNN	Mạng nơ-ron tích chập
DSC	Độ đo Dice
IoU	Độ đo giao trên hợp (Intersection over Union)
MAE	Mô hình Masked Autoencoders
RA	Chú ý ngược
RNN	Mạng nơ-ron hồi quy

# CHƯƠNG 1. GIỚI THIỆU ĐỀ TÀI

## 1.1 Đặt vấn đề

Trong lĩnh vực y học, việc chẩn đoán chính xác các bệnh lý đóng vai trò quan trọng, không chỉ giúp xác định đúng bệnh mà còn hỗ trợ trong việc điều trị kịp thời và hiệu quả. Nội soi y tế là một trong những phương pháp chẩn đoán hình ảnh phổ biến trong lĩnh vực y học, giúp bác sĩ quan sát trực tiếp các bệnh lý trên niêm mạc của các cơ quan trong cơ thể. Tuy nhiên, việc chẩn đoán bệnh qua hình ảnh nội soi vẫn còn phụ thuộc vào kinh nghiệm và kiến thức chuyên môn của bác sĩ, có thể dẫn đến sai sót hoặc bỏ sót bệnh lý. Theo một nghiên cứu về việc chẩn đoán bệnh thông qua hình ảnh nội soi, tỉ lệ bỏ sót bệnh nhân về các bệnh lý ung thư đường tiêu hóa đạt 9.8% [6]. Theo Tổ chức Ung thư toàn cầu (GLOBOCAN) năm 2020, Việt Nam mỗi năm có hơn 200.000 ca mắc mới và hơn 100.000 ca tử vong do bệnh ung thư, trong đó ung thư đường tiêu hóa chiếm hơn 30%. Tuy nhiên, các bệnh ung thư đường tiêu hóa có thể phòng ngừa và chữa trị hiệu quả nếu được phát hiện sớm. Vì vậy, việc áp dụng các thành tựu công nghệ như học máy vào việc chẩn đoán bệnh qua hình ảnh nội soi có thể giúp cải thiện chính xác và hiệu quả của quá trình chẩn đoán bệnh.

Các mô hình học máy có thể nhận diện các hình ảnh nội soi tiêu hóa có vùng tổn thương mà mắt thường có thể bỏ sót, giúp tự động phân đoạn và nhận dạng các vùng tổn thương trong hình ảnh nội soi, từ đó giúp giảm tải công việc cho các bác sĩ, nâng cao độ chính xác của chẩn đoán, cung cấp công cụ hỗ trợ đắc lực cho các bác sĩ trong quá trình chẩn đoán. Một số ứng dụng của AI trong chẩn đoán bệnh qua hình ảnh nội soi đã được nghiên cứu và phát triển, như việc chẩn đoán bệnh viêm loét hành tá tràng, viêm loét dạ dày, hay bệnh ung thư đường tiêu hóa.

Tuy nhiên, việc áp dụng học máy để nhận diện tổn thương trong hình ảnh y tế còn gặp nhiều thách thức. Để các mô hình hoạt động hiệu quả trong việc phát hiện chính xác các vùng tổn thương trong hình ảnh nội soi, không được bỏ sót thì cần phải có một lượng lớn dữ liệu hình ảnh nội soi đã được gắn nhãn. Tuy nhiên, việc thu thập và gắn nhãn dữ liệu cho hình ảnh nội soi y tế là một quá trình tốn kém về thời gian và công sức, đòi hỏi chuyên môn và kinh nghiệm cao từ các bác sĩ chuyên khoa.

Do đó, phương pháp học máy tự giám sát nổi lên như một giải pháp tiềm năng giúp khắc phục những khó khăn này. Thay vì phụ thuộc hoàn toàn vào dữ liệu gắn nhãn, phương pháp tự giám sát sử dụng các kỹ thuật học để khai thác thông tin từ dữ liệu chưa gắn nhãn. Cụ thể, mô hình tự giám sát có thể học các đặc trưng từ

dữ liệu hình ảnh thông qua các nhiệm vụ giả định (pretext task), chẳng hạn như tái tạo lại hình ảnh từ những phần bị che khuất hoặc dự đoán vị trí của các mảnh hình ảnh được xáo trộn. Những nhiệm vụ này không cần gắn nhãn dữ liệu mà vẫn có thể giúp mô hình học được các đặc trưng quan trọng của hình ảnh.

Sau khi được huấn luyện bằng phương pháp tự giám sát, mô hình có thể áp dụng vào các nhiệm vụ mục tiêu như nhận diện và phân đoạn tổn thương với một lượng nhỏ dữ liệu gắn nhãn. Điều này không chỉ giúp tiết kiệm thời gian và công sức trong quá trình thu thập và gắn nhãn dữ liệu mà còn giúp mô hình đạt hiệu suất cao hơn do đã được huấn luyện trên một lượng lớn dữ liệu chưa gắn nhãn.

Hơn nữa, phương pháp tự giám sát còn giúp tăng cường khả năng tổng quát hóa của mô hình, làm cho mô hình hoạt động tốt hơn trên các bộ dữ liệu mới mà không cần phải điều chỉnh nhiều. Điều này đặc biệt quan trọng trong y học, nơi mà sự đa dạng về hình ảnh và bệnh lý rất lớn. Nhờ khả năng học sâu từ dữ liệu đa dạng mà không cần gắn nhãn, các mô hình tự giám sát có thể phát hiện các đặc điểm tiềm ẩn và tinh vi trong hình ảnh nội soi mà có thể bị bỏ sót bởi con người hoặc các phương pháp học máy truyền thống.

Trong bối cảnh này, việc nghiên cứu và phát triển các mô hình học máy tự giám sát để huấn luyện mô hình nền tảng cho hình ảnh nội soi y tế là một hướng đi đầy tiềm năng và ý nghĩa. Nó không chỉ mở ra cơ hội cải tiến chất lượng chẩn đoán và điều trị trong y học mà còn góp phần giảm tải công việc cho các bác sĩ, tối ưu hóa quy trình chẩn đoán, và mang lại lợi ích to lớn cho sức khỏe cộng đồng.

## 1.2 Phạm vi của đồ án

Đồ án này tập trung vào việc sử dụng phương pháp học tự giám sát nhằm đào tạo ra một mô hình nền tảng cho hình ảnh nội soi y tế, tận dụng lượng lớn dữ liệu hình ảnh nội soi y tế không có nhãn trong y tế cho quá trình tiền huấn luyện và chỉ cần sử dụng một lượng nhỏ hình ảnh được gán nhãn để đào tạo mô hình cho các nhiệm vụ mục tiêu.

Mô hình được sử dụng để học từ các hình ảnh nội soi, một lĩnh vực đặc thù trong y tế với lượng dữ liệu phong phú nhưng phần lớn là những hình ảnh không có nhãn. Bằng cách tự đào tạo mô hình bằng phương pháp học tự giám sát, mô hình có thể học các đặc trưng từ dữ liệu không có nhãn thông qua các nhiệm vụ giả định, từ đó mô hình có thể khai quát hóa mãnh mẽ những đặc trưng của dữ liệu và trở thành một xương sống mạnh mẽ được sử dụng để tinh chỉnh với một tập dữ liệu nhỏ hơn đã có nhãn. Phương pháp này không chỉ tiết kiệm chi phí và thời gian gán nhãn dữ liệu vì không mất công sức vào việc gán nhãn cho dữ liệu trong quá trình tiền huấn luyện như các mô hình giám sát truyền thống trước đây khi mà các mô

hình đó không thể khám phá được hết tiềm năng to lớn của những dữ liệu không có nhãn. Không những vậy, mô hình được học tự giám sát còn cải thiện độ chính xác và nâng cao hiệu suất trong các nhiệm vụ mục tiêu.

### 1.3 Đóng góp của đồ án

Đồ án này có 3 đóng góp chính như sau:

1. Đồ án áp dụng kỹ thuật học tự giám sát để huấn luyện mô hình nền tảng tiên phong cho hình ảnh nội soi tiêu hoá nhằm tận dụng lượng lớn dữ liệu không có nhãn trong các trường hợp khác nhau. Kỹ thuật này giúp giảm thiểu sự phụ thuộc vào dữ liệu có nhãn, vốn rất tốn kém và khó thu thập trong lĩnh vực y tế.
2. Đề xuất một mô hình đa nhiệm cho hình ảnh nội soi y tế, cho phép mô hình có thể hoạt động trên nhiều nhiệm vụ khác nhau như phân loại, phân vùng tổn thương.
3. Thủ nghiệm trên các bộ dữ liệu khác nhau và từ kết quả thử nghiệm cho thấy, trong giai đoạn tinh chỉnh mô hình với nhiệm vụ mục tiêu trên cùng một lượng dữ liệu ảnh nội soi có nhãn, mô hình cơ sở được **tiếp tục** tiền huấn luyện bằng học tự giám sát trên số ít kỉ nguyên cho hiệu suất vượt trội hơn so với mô hình được huấn luyện từ đầu với số kỉ nguyên bằng số kỉ nguyên khi huấn luyện mô hình cơ sở trên ImageNet-1k [7]. Đồng thời, mô hình được tiếp tục tiền huấn luyện cũng cho hiệu suất cao hơn so với mô hình cơ sở.

### 1.4 Bô cục đồ án

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

Chương 2 trình bày về khái niệm, cơ sở lý thuyết liên quan đến học máy, trí tuệ nhân tạo, học sâu và các mô hình liên quan đến phân vùng, đa nhiệm.

Chương 3 đề xuất các mô hình giải quyết các nhiệm vụ liên quan hình ảnh nội soi tiêu hóa.

Chương 4 trình bày nội dung thử nghiệm và thảo luận về những kết quả thu được.

Chương 5 kết luận và hướng phát triển trong tương lai.

## CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT

Trong chương này, sẽ trình bày các kiến thức nền tảng về học máy và học sâu, các phương pháp học máy, các mô hình học sâu, học đa nhiệm, v.v. Đồng thời, chúng tôi cũng trình bày về các nghiên cứu liên quan đến phân đoạn hình ảnh nội soi y tế và phân tích rõ những ưu nhược điểm của chúng. Từ đó, nêu bật lên động lực để thực hiện nghiên cứu của đồ án này.

### 2.1 Khái niệm về học máy và học sâu

Học máy là một lĩnh vực của trí tuệ nhân tạo, tập trung vào việc phát triển các kỹ thuật để máy tính có thể học từ dữ liệu mà không cần phải được lập trình cụ thể. Học máy được chia thành nhiều loại khác nhau dựa trên cách mà dữ liệu được cung cấp cho mô hình học máy.

Học sâu là một lĩnh vực con của học máy, nó tập trung vào việc xây dựng và huấn luyện các mô hình mạng nơ-ron nhân tạo để giải quyết các bài toán phức tạp. Một số loại mô hình học sâu phổ biến hiện nay bao gồm mạng nơ-ron tích chập (Convolutional Neural Network - CNN) [5], mô hình Transformer [8], mô hình Vision Transformers [9], ...

### 2.2 Các phương pháp học máy

**Học có giám sát** là một kỹ thuật mà trong đó máy tính được học từ dữ liệu đã được gán nhãn để phát triển các thuật toán có khả năng phân loại hoặc dự đoán kết quả một cách chính xác. Ở kỹ thuật này, các giá trị đầu vào và đầu ra đã được biết trước và thuật toán học máy sẽ học hàm ánh xạ từ đầu vào đến đầu ra. Về mặt toán học, với  $Y$  là đầu ra và  $X$  là đầu vào, các thuật toán học máy cố gắng tìm ra hàm ánh xạ tốt nhất  $f$  sao cho  $Y = f(X)$ . Nếu quan sát kỹ, chúng ta thấy việc học diễn ra giống như một người giám sát nào đó giám sát quá trình học tập. Chúng ta đã biết câu trả lời nên thuật toán cố gắng ánh xạ hàm sao cho đầu ra dự đoán phải gần với đầu ra thực tế.

Giả sử mô hình đã học được hàm ánh xạ  $f$  dự đoán các giá trị  $Y'$  cho mỗi  $X$  được truyền vào mô hình. Khi sự khác biệt giữa dự đoán ( $Y'$ ) và thực tế ( $Y$ ) giảm xuống dưới một ngưỡng nhất định (nói một cách đơn giản, sự sai khác trở nên không đáng kể), quá trình học sẽ dừng lại.

Học có giám sát có thể được chia thành 2 loại là phân loại và hồi quy. Các ứng dụng thực tế của học có giám sát như dự đoán rủi ro, phát hiện đối tượng, phân loại hình ảnh, phát hiện gian lận, hệ thống đề xuất, dự đoán chuỗi thời gian, ...

**Học không giám sát** là một kỹ thuật mà trong đó chúng ta chỉ có dữ liệu đầu

vào để cung cấp cho mô hình nhưng không có dữ liệu đầu ra tương ứng. Học không giám sát được thiết kế nhằm cố gắng phát hiện những điểm tương đồng của dữ liệu để lọc, chia dữ liệu vào các danh mục hoặc tìm những điểm bất thường trong dữ liệu. Học không giám sát thường được sử dụng trong các bài toán phân cụm, phân tích liên kết, giảm chiều dữ liệu và được sử dụng trong Mạng sáng tạo đối nghịch (GAN) [10].

**Học bán giám sát** là một kỹ thuật học máy sử dụng một lượng nhỏ dữ liệu gắn nhãn và một lượng lớn dữ liệu chưa gắn nhãn để huấn luyện mô hình. Dữ liệu được gán nhãn sẽ được dùng để huấn luyện cơ bản một mô hình. Sau đó mô hình đã qua huấn luyện cơ bản này sẽ được sử dụng để gắn nhãn dữ liệu chưa gắn nhãn – một quy trình gọi là “gắn nhãn giả”. Mô hình sau đó sẽ được huấn luyện bằng dữ liệu gắn nhãn và dữ liệu gắn nhãn giả. Quá trình này sẽ được lặp lại nhiều lần cho đến khi mô hình đạt được độ chính xác mong muốn. Do mô hình sử dụng được gán nhãn một lượng nhỏ dữ liệu được gán nhãn và một lượng lớn dữ liệu chưa gắn nhãn nên học bán giám sát giúp giảm thiểu chi phí gắn nhãn thủ công và giảm thời gian chuẩn bị dữ liệu. Vì dữ liệu không được gắn nhãn rất phong phú, dễ thu thập và không tốn kém nên học bán giám sát được tìm thấy nhiều ứng dụng trong khi độ chính xác của kết quả không bị ảnh hưởng.

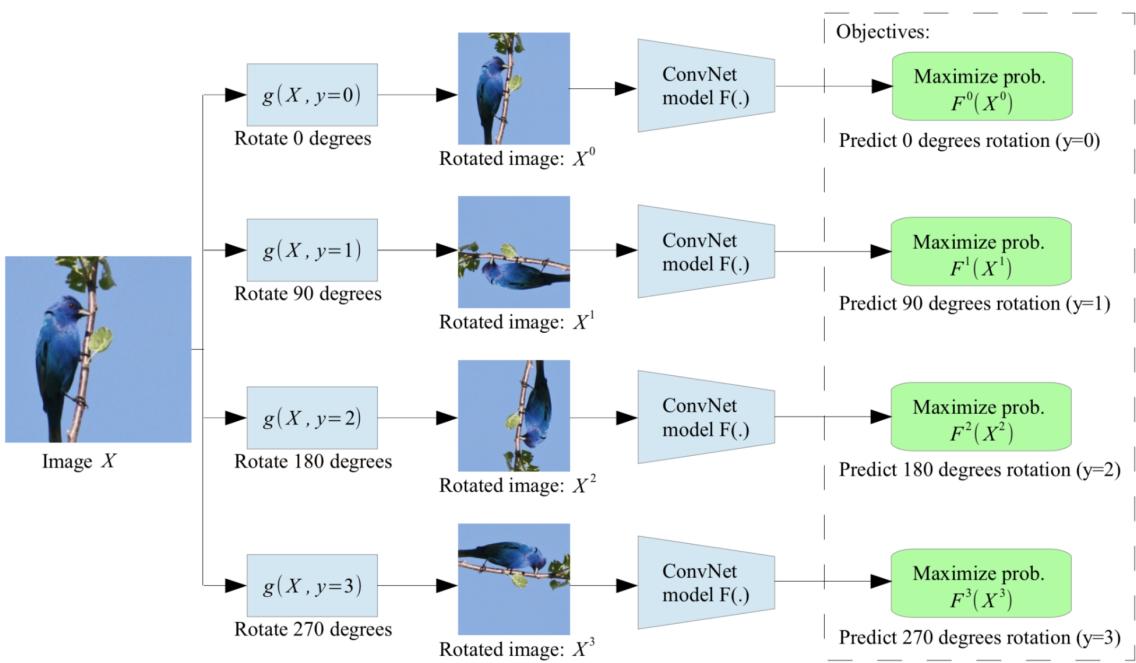
**Học tăng cường** là một loại kỹ thuật máy học mà một mô hình tự động tìm hiểu và cải thiện hành vi thông qua tương tác với môi trường. Quy trình này dựa trên nguyên tắc học từ phản hồi và thưởng để tối đa hóa một hàm phần thưởng được xác định trước. Lấy một ví dụ đơn giản như việc ta dạy một chú chó nhặt càành cây cho chúng ta sau khi ta đã ném nó đi. Mỗi khi chú chó nhặt càành cây cho ta, ta sẽ thưởng cho nó một khúc xương. Dần dần, chú chó hiểu được việc nhặt càành cây càng nhanh càng tốt thì chú chó sẽ được thưởng nhanh hơn. Các ứng dụng phổ biến có thể nhìn thấy trong cuộc sống hằng ngày của chúng ta như ứng dụng trong xe tự lái, ứng dụng trong tự động hóa công nghiệp, robotics, ...

**Học tự giám sát** là một kỹ thuật học máy mà mô hình có thể tự đào tạo mà không cần dữ liệu có nhãn. Các mô hình này được tiền huấn luyện trên một số nhiệm vụ giả định trên dữ liệu không có nhãn, chẳng hạn như dự đoán các pixel xung quanh một pixel nhất định hoặc tái tạo dữ liệu ban đầu. Sau đó, mô hình sẽ được đào tạo lại bằng cách sử dụng phương pháp học có giám sát (hoặc không giám sát) trên một tập dữ liệu (nhỏ) khác để giải quyết nhiệm vụ mục tiêu ban đầu. Hình 2.1 là một ví dụ về phương pháp huấn luyện tự giám sát.

Trong phương pháp học có giám sát yêu cầu lượng lớn dữ liệu có nhãn để huấn

---

<sup>1</sup><https://lilianweng.github.io/posts/2019-11-10-self-supervised/>



**Hình 2.1:** Huấn luyện mô hình tự giám sát với nhiệm vụ giả định là các phép xoay ảnh.<sup>1</sup>

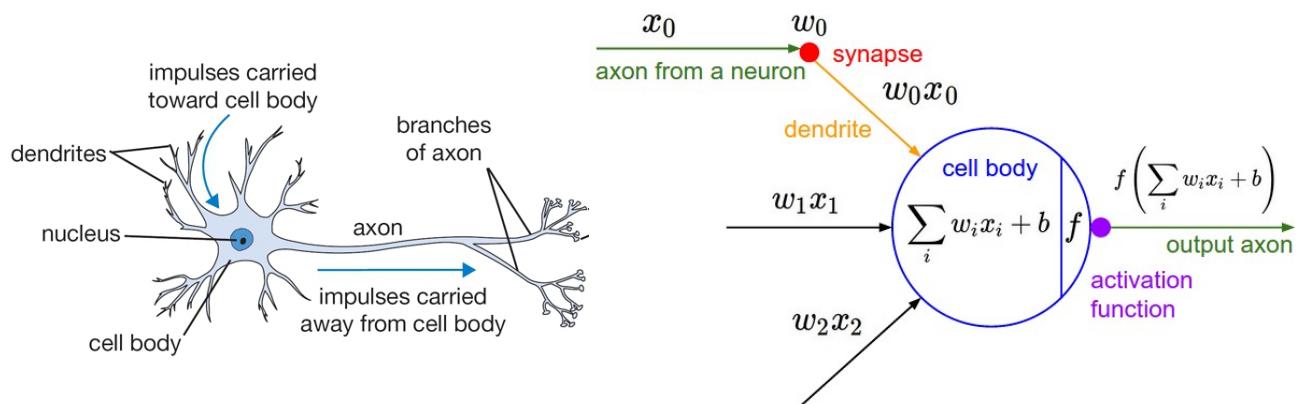
luyện mô hình trong cả hai quá trình tiền huấn luyện và tinh chỉnh. Tuy nhiên để gán nhãn một lượng lớn dữ liệu như thế cần phải thực hiện thủ hiện gây ra sự lãng phí tiền bạc và thời gian. Học tự giám sát nổi lên như một giải pháp tiềm năng nhờ vào việc chỉ cần gán nhãn một lượng nhỏ dữ liệu trong quá trình tinh chỉnh mà không cần gán nhãn cho quá trình tiền huấn luyện. Việc gán nhãn trong quá trình tiền huấn luyện của phương pháp này được thực hiện bằng các nhiệm vụ giả định. Nhờ vào ưu điểm của phương pháp này vì không phát sinh chi phí gán nhãn lượng lớn dữ liệu nên phương pháp học tự giám sát đang được áp dụng rộng rãi trong mô hình ngôn ngữ lớn, mô hình nền tảng, ...

Việc tự động tạo nhãn giả cho hình ảnh trong lĩnh vực thị giác máy tính như việc tăng cường dữ liệu hình ảnh bằng cách biến đổi như cắt ngẫu nhiên hoặc biến dạng màu sắc, sau đó sử dụng chúng cho các nhiệm vụ giả định bằng cách so sánh giữa các hình ảnh với nhau. Trong tác vụ này, hình ảnh gốc được gọi là "mỏ neo", hình ảnh mở rộng cùng nguồn gốc với mỏ neo được gọi là "mẫu dương tính", và hình ảnh mở rộng khác nguồn gốc với mỏ neo được gọi là "mẫu âm tính". Quá trình học tập sẽ tối đa hóa khoảng cách giữa "mỏ neo" và "mẫu dương tính", đồng thời tối thiểu hóa khoảng cách giữa "mỏ neo" và "mẫu âm tính". Đây là một phương pháp học tập tiêu biểu của học tự giám sát có tên gọi là học tương phản, ví dụ như SimCLR [11]. Ngoài ra, còn có nhiều phương pháp khác đã được đề xuất như mô hình hóa che hình ảnh (Masked Image Modeling), ví dụ như SimMIM [11], BEiT [1], iBot [2], ...

## 2.3 Các mô hình học sâu

### 2.3.1 Mạng nơ-ron nhân tạo

Học sâu, một nhánh đặc biệt của học máy, đã nổi lên như một công cụ mạnh mẽ để giải quyết những bài toán phức tạp nhờ khả năng tự động học ra các đặc trưng từ dữ liệu thô. Học sâu sử dụng các mạng nơ-ron nhân tạo nhiều lớp để mô phỏng cách hoạt động của não người như Hình 2.2. Nhờ vào cấu trúc đặc biệt này, các mô hình mạng nơ-ron có thể học các đặc trưng ở nhiều mức độ trừu tượng khác nhau. Mạng nơ-ron nhân tạo có cấu tạo có thể có nhiều lớp, mỗi lớp chứa nhiều nơ-ron, các lớp được kết nối với nhau bằng các trọng số. Mỗi nơ-ron trong một lớp sẽ nhận đầu vào từ tất cả các nơ-ron trong lớp trước đó, tính toán thông qua trọng số và hàm kích hoạt để tạo ra đầu ra để dự đoán hoặc phân loại dữ liệu. Mô hình sẽ học cách điều chỉnh các trọng số này thông qua quá trình lan truyền ngược và hàm tổn thất, giúp mô hình học được cách biểu diễn dữ liệu.



**Hình 2.2:** Mạng học sâu nơ-ron được lấy cảm hứng từ não người.<sup>2</sup>

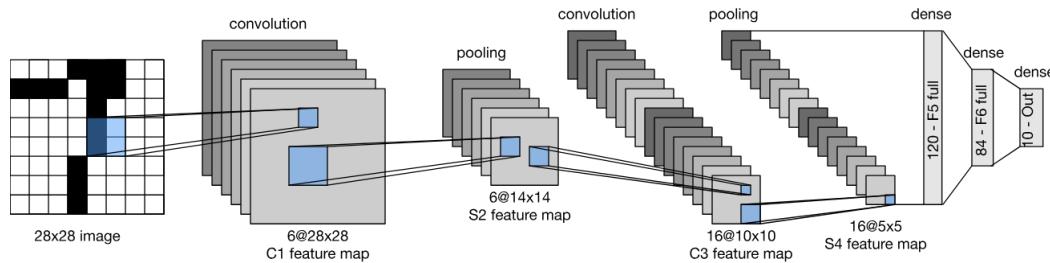
### 2.3.2 Mạng nơ-ron tích chập

Sau sự thành công của mạng nơ-ron nhân tạo (ANN) [12] trong việc giải quyết nhiều bài toán khác nhau, các nhà nghiên cứu nhận thấy rằng ANN vẫn còn một số hạn chế khi xử lý các dữ liệu có cấu trúc không gian, chẳng hạn như hình ảnh. Nếu sử dụng các mạng nơ-ron nhiều lớp (Multi-layer Perceptron- MLP) [12] cho việc phân loại hình ảnh, chúng ta cần phải làm phẳng hình ảnh thành một vector dữ liệu trước khi truyền chúng qua mạng, điều này dẫn đến việc mất mát thông tin không gian. Có một điều rất quan trọng là các điểm ảnh liền kề nhau có mối quan hệ với nhau, ta cần có một mô hình tận dụng được điều này để xây dựng một mô hình hiệu quả hơn cho việc học từ dữ liệu ảnh. Để giải quyết vấn đề này, các nhà nghiên cứu đã đề xuất một kiến trúc mạng nơ-ron mới, được gọi là mạng nơ-ron tích chập (Convolutional Neural Network - CNN) [5][12]. Kiến trúc và cách hoạt

<sup>2</sup><https://cs231n.github.io/neural-networks-1>

động của mạng CNN được mô tả ở Hình 2.2.

Sau khi được giới thiệu, CNN đã chứng minh được hiệu quả trong việc xử lý dữ liệu ảnh và nhanh chóng trở thành một công cụ quan trọng trong nhiều ứng dụng. Hiện nay, CNN xuất hiện khắp mọi ngõ ngách của lĩnh vực thị giác máy tính và đã trở thành kiến trúc chính mà ai xây dựng mô hình sẽ dựa trên nó.



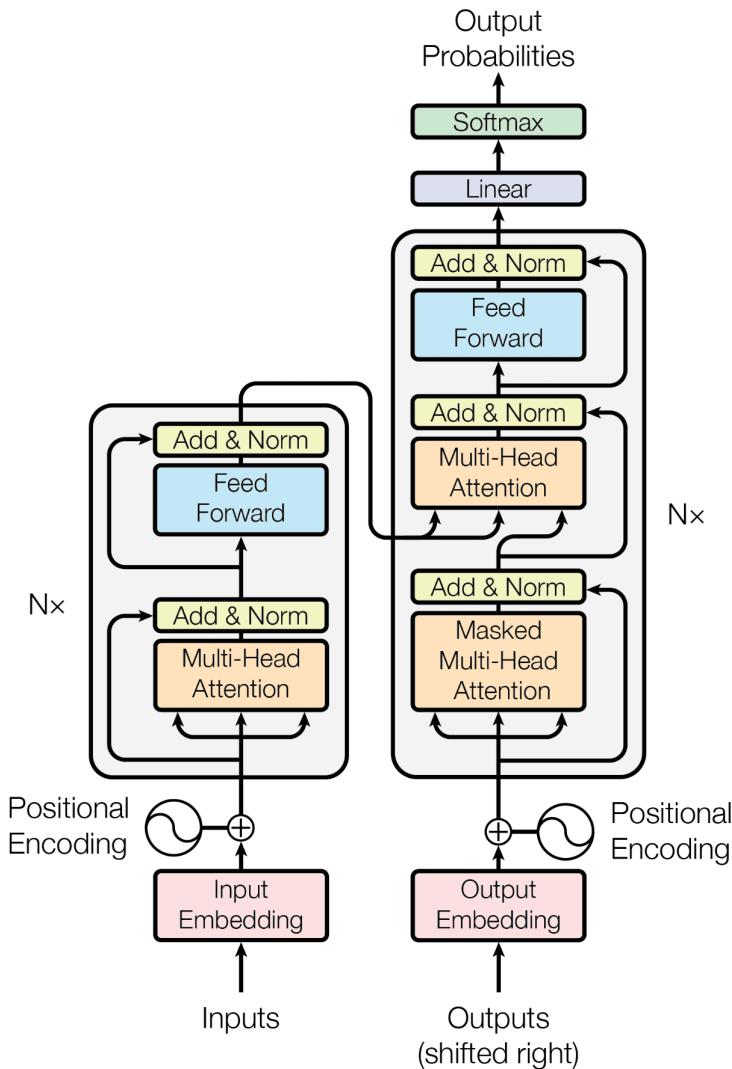
**Hình 2.3:** Kiến trúc mạng nơ-ron tích chập - CNN [12]

Hoạt động của CNN dựa trên ba loại lớp chính: lớp tích chập (convolutional layer), lớp gộp (pooling layer), và lớp kết nối đầy đủ (fully connected layer). Lớp tích chập sử dụng các bộ lọc (filter) để quét qua đầu vào, tạo ra các bản đồ đặc trưng (feature maps) bằng cách thực hiện phép tích chập (convolution) giữa bộ lọc và các vùng nhỏ của đầu vào. Mỗi bộ lọc học cách phát hiện các đặc trưng cụ thể như cạnh, góc hoặc các hình dạng phức tạp hơn trong hình ảnh. Kích thước của bộ lọc và bước dịch chuyển (stride) quyết định cách mà bộ lọc quét qua đầu vào và kích thước của bản đồ đặc trưng kết quả. Lớp gộp, thường là gộp tối đa (max pooling), được sử dụng để giảm kích thước không gian của bản đồ đặc trưng, giúp giảm số lượng tham số và tính toán trong mạng, cũng như giảm thiểu nguy cơ quá khớp (overfitting). Trong quá trình này, giá trị lớn nhất trong mỗi vùng nhỏ của bản đồ đặc trưng được chọn để tạo ra bản đồ đặc trưng đã gộp. Cuối cùng, lớp kết nối đầy đủ tương tự như mạng nơ-ron truyền thống. Các lớp này thường được sử dụng ở phần cuối của mạng để thực hiện các tác vụ phân loại.

### 2.3.3 Transformers

Transformer là một cấu trúc mạng nơ-ron mạnh mẽ trong lĩnh vực học máy, được giới thiệu lần đầu trong bài báo "**Attention is all you need**" [8] của Vaswani và đồng nghiệp vào năm 2017. Từ khi xuất hiện, nó đã tạo ra một cuộc cách mạng trong việc xử lý ngôn ngữ tự nhiên, thay thế cho các mô hình RNN và LSTM từng phổ biến trước đó. Transformer có cấu trúc bộ mã hóa - bộ giải mã. Bộ mã hóa nhận một chuỗi các ký hiệu ( $x_1, x_2, \dots, x_n$ ) và chuyển đổi nó thành một biểu diễn

liên tục ( $z_1, z_2, \dots, z_n$ ) để nắm bắt toàn bộ thông tin đã học từ đầu vào. Sử dụng biểu diễn này, bộ giải mã tạo ra một chuỗi đầu ra ( $y_1, y_2, \dots, y_m$ ) bằng cách sinh ra từng ký hiệu. Mô hình hoạt động theo kiểu tự hồi quy, trong đó mô hình sử dụng các ký hiệu đã sinh ra trước đó làm đầu vào bổ sung ở mỗi bước để sinh ra ký hiệu tiếp theo. Đối với cả bộ mã hóa và bộ giải mã, Transformer đều sử dụng cơ chế tự chú ý (self-attention).



**Hình 2.4:** Kiến trúc Transformers [8]

Kiến trúc của Transformers (xem Hình 2.4) có các thành phần như sau:

**Hàm chú ý** (Scaled dot-product Attention) ánh xạ một truy vấn (query) và một tập hợp các cặp khóa-giá trị (key-value) thành một đầu ra. Trong đó, đầu vào (input) bao gồm các truy vấn và khóa có chiều  $d_k$  và các giá trị có chiều  $d_v$ . Với  $Q$  là ma trận truy vấn,  $K$  là ma trận khóa,  $V$  là ma trận giá trị, hàm chú ý được tính như sau:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (2.1)$$

Các tác giả chỉ ra rằng đối với các giá trị lớn của  $d_k$ , tích vô hướng trở nên rất lớn nên sẽ đẩy hàm softmax vào các vùng mà nó có gradient cực kỳ nhỏ. Do đó, trong công thức, tích vô hướng được chia tỷ lệ bởi  $\sqrt{d_k}$  để giảm độ lớn của tích vô hướng và giữ cho gradient ổn định.

**Chú ý đa đầu** (Multi-head Attention) cho phép mô hình chú ý nắm bắt thông tin từ nhiều không gian con khác nhau. Mỗi head sẽ học một cách chú ý khác nhau từ các cặp khóa-giá trị. Cụ thể, với  $h$  head, mỗi head sẽ học một ma trận truy vấn  $Q_i$ , khóa  $K_i$  và giá trị  $V_i$ . Các ma trận truy vấn, khóa và giá trị được tạo ra từ truy vấn, khóa và giá trị ban đầu bằng cách nhân tương ứng với các ma trận trọng số  $W_i^Q$ ,  $W_i^K$  và  $W_i^V$ . Công thức của multi-head attention được tính như sau:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2.2)$$

trong đó  $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$  và  $W^O$  là ma trận trọng số đầu ra.

**Mã hóa vị trí** (Position Encoding) được thêm vào chuỗi đầu vào để mô hình có thể học được thông tin về vị trí của từng từ trong chuỗi. Cụ thể, với mỗi từ tại vị trí  $i$ , vector biểu diễn của từ sẽ được cộng với một ma trận biểu diễn vị trí  $PE_i$ :

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (2.3)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{\text{model}}}}\right) \quad (2.4)$$

**Mạng chuyển tiếp nguồn cấp dữ liệu theo vị trí** (Position-wise Feed-Forward Networks) là một mạng nơ-ron với hai lớp tuyến tính kết nối đầy đủ. Với một chuỗi đầu vào có chiều  $d_{\text{model}}$ , mạng sẽ ánh xạ mỗi từ tại vị trí  $i$  thành một vector có chiều  $d_{\text{ff}}$  bằng cách sử dụng hai ma trận trọng số  $W_1$  và  $W_2$  và hàm kích hoạt ReLU:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2.5)$$

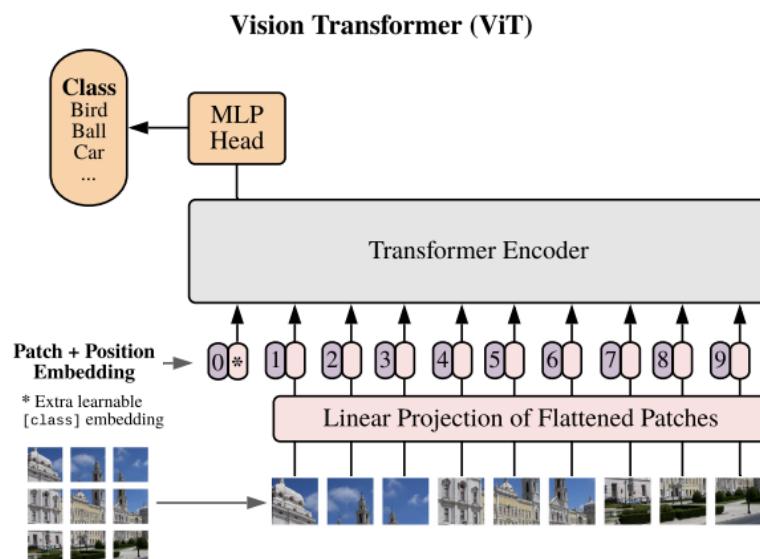
**Kết nối tắt** (Residual Connection [13]) và **Chuẩn hóa theo Layer** (Layer Normalization [14]) được sử dụng để giải quyết vấn đề biến mất gradient [15] trong quá trình huấn luyện mạng sâu. Cụ thể, các kết nối tắt được thêm vào mỗi lớp của mô hình và sau đó được chuẩn hóa bằng cách sử dụng chuẩn hóa theo lớp (layer normalization):

$$\text{LayerNorm}(x + \text{Sublayer}(x)) \quad (2.6)$$

### 2.3.4 Vision Transformers

Vision Transformer (ViT) là mô hình học sâu tiên tiến được giới thiệu lần đầu tiên trong bài báo "**An image is worth 16x16 words: Transformers for image recognition at scale**" bởi Dosovitskiy và cộng sự vào năm 2020 [9]. Mô hình này đánh dấu một bước ngoặt trong lĩnh vực thị giác máy tính khi đưa ra một cách tiếp cận mới mẻ và hiệu quả cho các tác vụ nhận dạng hình ảnh.

Điểm khác biệt chính của ViT so với các mô hình truyền thống như mạng nơ-ron tích chập (CNN) [5] là nó sử dụng kiến trúc Transformers [8] thay vì dựa vào các thao tác tích chập. Transformer, vốn nổi tiếng trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), được áp dụng cho ViT [9] để xử lý hình ảnh bằng cách chia nhỏ hình ảnh thành các "patch" và sử dụng cơ chế "self-attention" để học hỏi mối quan hệ giữa các patch này.



**Hình 2.5:** Kiến trúc Vision Transformer [9]

Lấy cảm hứng từ kiến trúc của Transformers, kiến trúc của Transformers (ViT) [9] trong lĩnh vực thị giác máy tính chỉ sử dụng bộ mã hóa (encoder) của Transformers, bỏ qua bộ giải mã (decoder), thêm một lớp phân loại ở cuối cùng để dự đoán nhãn của hình ảnh và thêm một lớp Patch Embedding để chuyển đổi hình ảnh thành dạng chuỗi các vector giúp Transformers [8] có thể xử lý hình ảnh như một chuỗi dữ liệu, tận dụng thế mạnh của kiến trúc Transformers [8] trong việc học các mối quan hệ giữa các phần tử trong chuỗi. Hình 2.5 minh họa kiến trúc của ViT [9].

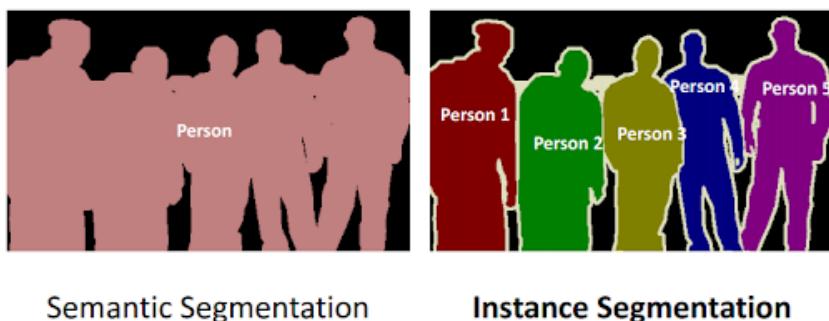
ViT [9] đã chứng minh khả năng vượt trội so với các mô hình CNN [5] trên nhiều điểm chuẩn (benchmark) về nhận dạng hình ảnh, đạt được độ chính xác cao

hơn với hiệu quả tính toán tốt hơn. Ưu điểm này của ViT [9] đến từ khả năng học tập các mối quan hệ tầm xa trong hình ảnh một cách hiệu quả, điều mà các mô hình CNN [5] gặp nhiều khó khăn. Sự ra đời của ViT [9] đã mở ra hướng nghiên cứu mới cho lĩnh vực thị giác máy tính, thúc đẩy sự phát triển của các mô hình học sâu hiệu quả và chính xác hơn cho các ứng dụng nhận dạng hình ảnh, phân loại ảnh, và xử lý ảnh.

## 2.4 Bài toán phân vùng ảnh y tế và các mô hình học sâu

### 2.4.1 Bài toán phân vùng ảnh

là một trong những nhiệm vụ quan trọng trong lĩnh vực xử lý ảnh và thị giác máy tính. Nó liên quan đến việc chia một bức ảnh thành các vùng khác nhau sao cho mỗi vùng đại diện cho một đối tượng hoặc một phần của đối tượng nào đó trong ảnh. Phân vùng ảnh có các loại như phân vùng ngữ nghĩa (semantic segmentation) và phân vùng đối tượng (instance segmentation). Phân vùng ngữ nghĩa có nhiệm vụ phân loại mỗi pixel trong ảnh vào một trong các lớp đối tượng đã biết, trong khi phân vùng đối tượng không chỉ gán nhãn cho mỗi pixel mà còn phân biệt từng đối tượng riêng lẻ trong cùng một lớp.



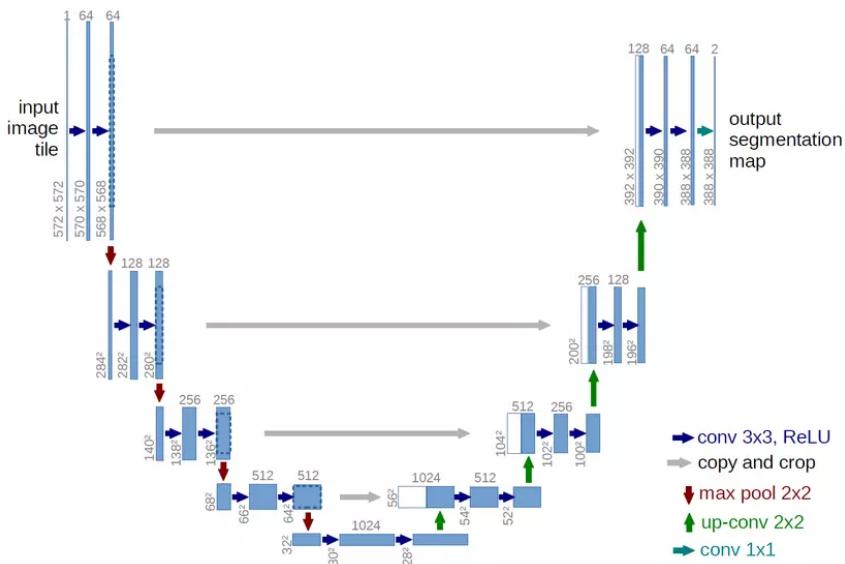
**Hình 2.6:** Phân vùng ngữ nghĩa và phân vùng đối tượng<sup>3</sup>

### 2.4.2 Giám sát sâu (Deep supervision)

### 2.4.3 U-Net

Mạng U-Net (viết tắt của "U-shaped Neural Network for Image Segmentation") là một kiến trúc mạng nơ-ron nhân tạo được phát triển ban đầu cho mục đích phân đoạn hình ảnh y tế. Được giới thiệu lần đầu tiên trong bài báo "U-Net: Convolutional Networks for Biomedical Image Segmentation" vào năm 2015 bởi Olaf Ronneberger và cộng sự [16]. Tuy nhiên, do hiệu quả vượt trội của nó, U-Net đã được ứng dụng rộng rãi trong nhiều lĩnh vực khác như xử lý ảnh vệ tinh, phân tích ảnh vi mô, và robot.

<sup>3</sup><https://folio3.ai/blog/semantic-segmentation-vs-instance-segmentation>



Hình 2.7: Kiến trúc UNet [16]

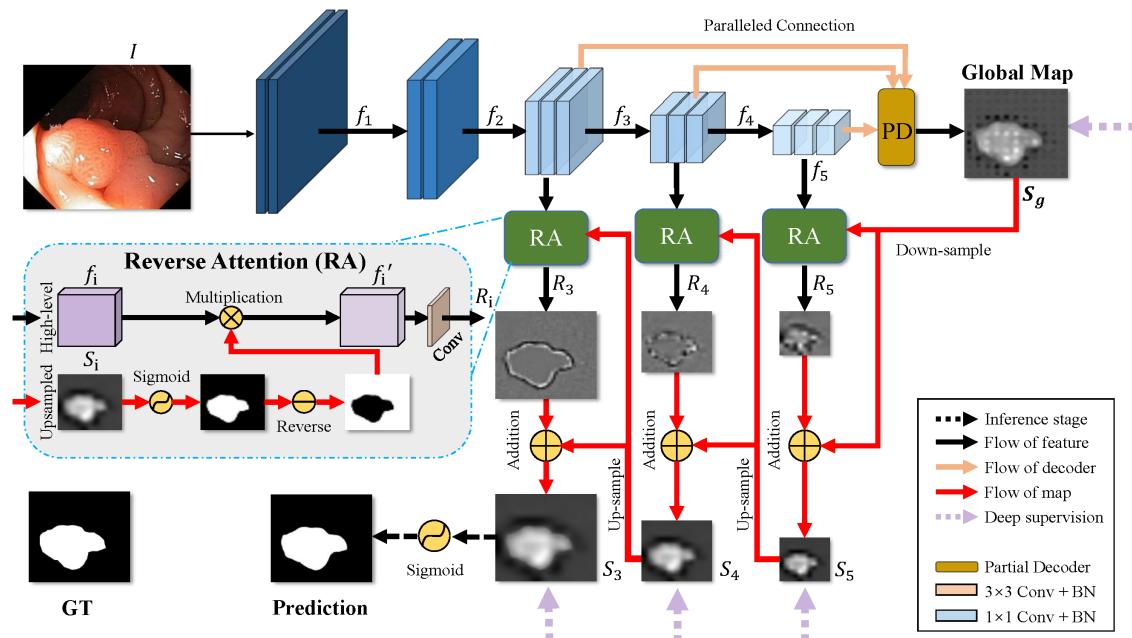
**Kiến trúc U-Net** có hình dạng chữ U đặc trưng, bao gồm hai phần chính:

- **Bộ mã hóa (Encoder):** Gồm các khối tích chập (convolutional block) được sắp xếp theo thứ tự giảm dần kích thước ảnh (downsampling). Mỗi khối tích chập bao gồm các lớp tích chập (convolution layer), lớp kích hoạt (activation layer) và lớp pooling (pooling layer). Quá trình mã hóa giúp trích xuất các đặc trưng cấp cao từ ảnh đầu vào.
- **Bộ giải mã (Decoder):** Gồm các khối upsampling được sắp xếp theo thứ tự tăng dần kích thước ảnh (upsampling). Mỗi khối upsampling bao gồm các lớp tích chập chuyển đổi (transposed convolutional layer), lớp kích hoạt và lớp nối (concatenation layer). Quá trình giải mã kết hợp các đặc trưng cấp cao được trích xuất từ encoder với các đặc trưng cấp thấp hơn để tạo ra bản đồ phân đoạn chi tiết.

#### 2.4.4 PraNet

PraNet là một mô hình phân vùng polyp trong hình ảnh nội soi được giới thiệu trong bài báo có tên "PraNet: Parallel Reverse Attention Network for Polyp Segmentation" vào năm 2020 với Deng-Ping Fan và cộng sự [17]. Bài báo này trình bày một phương pháp mới để phân đoạn polyp từ các hình ảnh nội soi. Bài báo đã nêu ra khó khăn khi phân vùng polyp vì sự đa dạng về kích thước, màu sắc và kết cấu của polyp, cùng với ranh giới không rõ ràng giữa polyp và mô xung quanh. Để giải quyết những vấn đề này, bài báo đề xuất mạng lưới chú ý ngược song song (Parallel Reverse Attention Network), sử dụng bộ giải mã bán phần song song (Parallel Partial Decoder) để tổng hợp các đặc trưng từ các lớp cao và tạo ra một bản đồ toàn cục. Các module chú ý ngược (Reserve Attention [18]) sau đó được sử dụng

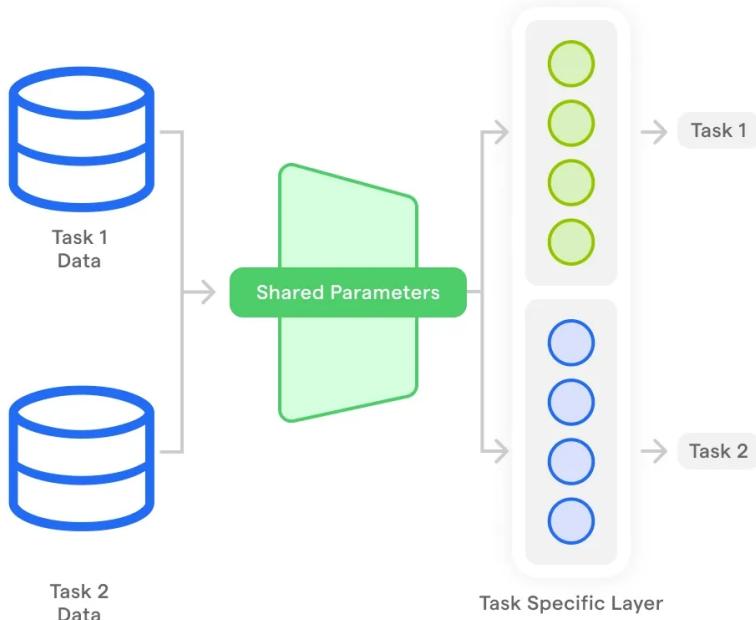
để khai thác các gợi ý về ranh giới, giúp cải thiện độ chính xác của phân đoạn. PraNet [17] có khả năng hiệu chỉnh các dự đoán sai lệch, nâng cao độ chính xác của phân đoạn, khả năng tổng quát tốt.



**Hình 2.8:** Kiến trúc của mô hình PraNet [17]

## 2.5 Mô hình đa nhiệm

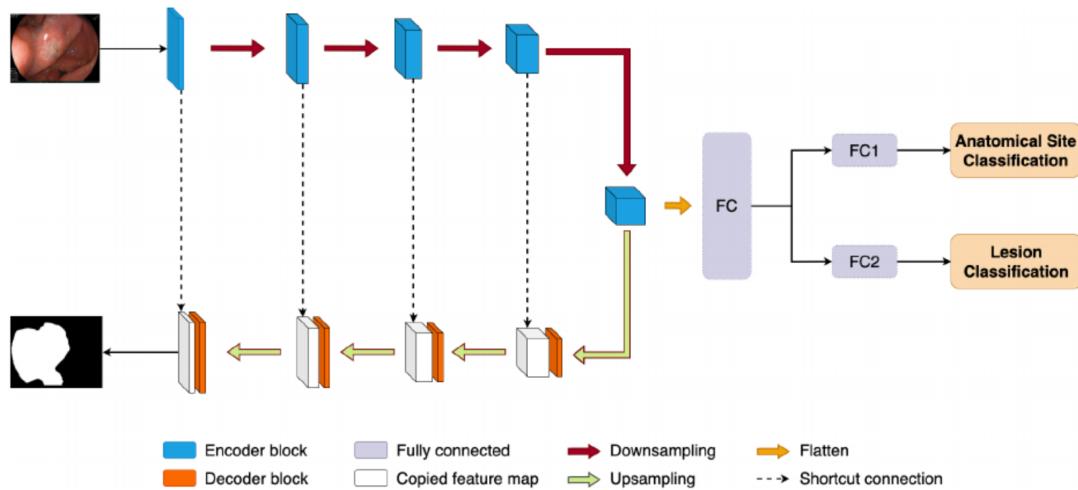
# Multitask Learning



**Hình 2.9:** Minh họa mô hình đa nhiệm <sup>4</sup>

Mô hình đa nhiệm (multitask model) là một loại mô hình học máy được thiết kế để giải quyết đồng thời nhiều nhiệm vụ khác nhau. Thay vì huấn luyện các mô hình riêng biệt cho từng nhiệm vụ, mô hình đa nhiệm học cách thực hiện nhiều nhiệm vụ trong cùng một kiến trúc, chia sẻ các tham số và thông tin giữa các nhiệm vụ. Điều này giúp cải thiện hiệu quả học tập và hiệu suất tổng thể của mô hình. Mô hình đa nhiệm thường bao gồm một xương sống (backbone) chung, trọng số được chia sẻ cho tất cả các nhiệm vụ, và các đầu (head) riêng biệt cho mỗi nhiệm vụ cụ thể (xem Hình 2.9). Xương sống (backbone) của mô hình học các đặc trưng chung có ích cho tất cả các nhiệm vụ, trong khi phần đầu (head) riêng biệt tinh chỉnh các đặc trưng này để phù hợp với từng nhiệm vụ cụ thể. Một trong những lợi ích chính của mô hình đa nhiệm là khả năng tổng quát hóa tốt hơn, vì mô hình học được từ nhiều nguồn dữ liệu và nhiệm vụ khác nhau, có khả năng áp dụng hiệu quả cho các tình huống thực tế hơn.

### EndoUNet



**Hình 2.10:** Kiến trúc EndoUNet [19]

Lấy cảm hứng từ mô hình đa nhiệm, mô hình EndoUNet [19] được thiết kế để giải quyết đồng thời hai nhiệm vụ phân đoạn và phân loại. Mô hình EndoUNet bao gồm một xương sống (backbone) chung, một phân bộ mã hóa chia sẻ giữa hai nhiệm vụ, và ba đầu giải mã (decoder) riêng biệt cho mỗi nhiệm vụ (xem Hình 2.10). EndoUNet có kiến trúc giống với mô hình U-Net [16], tuy nhiên để mô hình có thể thực hiện được đa nhiệm vụ, mô hình đã thêm ba đầu giải mã là các mạng nơ-ron tuyển tính lấy các đặc trưng thấp nhất của mạng U-Net để làm đầu vào. Mỗi đầu giải mã sẽ tinh chỉnh các đặc trưng này để phân loại ba nhiệm vụ tương ứng là phân loại HP, phân loại vị trí giải phẫu và phân loại loại bệnh. Mô hình EndoUNet được

<sup>4</sup><https://botpenguin.com/glossary/multi-task-learning>

## CHƯƠNG 2. NỀN TẢNG LÝ THUYẾT

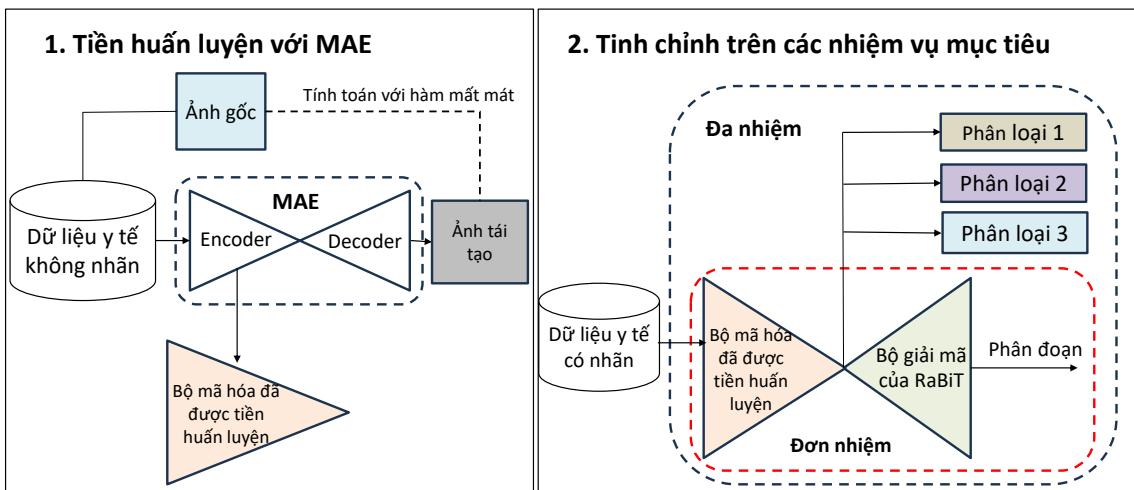
---

huấn luyện đồng thời trên cả hai nhiệm vụ, chia sẻ thông tin giữa phân đoạn ảnh và phân loại bệnh, giúp cải thiện hiệu suất tổng thể của mô hình.

## CHƯƠNG 3. PHƯƠNG PHÁP ĐỀ XUẤT

### 3.1 Tổng quan

Trong chương này, trình bày ngắn gọn việc tạo ra và sử dụng mô hình nền tảng cho hình ảnh nội soi y tế bằng phương pháp tự giám sát của mô hình MAE [4] trên một lượng dữ liệu hình ảnh nội soi y tế không có nhãn. Mô hình nền tảng được tạo ra trong giai đoạn tiền huấn luyện và được sử dụng trong giai đoạn tinh chỉnh trên các nhiệm vụ mục tiêu bằng cách chỉ lấy bộ mã hóa kết hợp với bộ giải mã khác để thực hiện các nhiệm vụ mục tiêu. Thiết kế tổng thể của chúng tôi được thể hiện trong Hình 3.1.



**Hình 3.1:** Thiết kế tổng quan sử dụng MAE [4] cho quá trình tiền huấn luyện và tinh chỉnh

Để mô hình Masked Autoencoders (MAE) [4] học được cách biểu diễn dữ liệu mạnh mẽ hơn trong giai đoạn tiền huấn luyện và đạt được hiệu suất cao hơn nữa các nhiệm vụ mục tiêu ở giai đoạn tinh chỉnh phía sau, chúng tôi đã thêm vào bộ mã hóa (encoder) của mô hình MAE [4] một số token đặc biệt bên cạnh token `cls` của Vision Transformer [9]. Trong giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu như phân vùng hình ảnh, chúng tôi sử dụng bộ giải mã mạnh mẽ của RaBiT [20] nhằm giúp mô hình đạt hiệu suất cao trên các nhiệm vụ mục tiêu. Bên cạnh mô hình đơn nhiệm, chúng tôi cũng đề xuất một mô hình đa nhiệm kết hợp cả nhiệm vụ phân vùng tổn thương và phân loại bệnh trong một mô hình duy nhất và mô hình đa nhiệm này có thể ứng dụng trong thực tế nhằm giúp đỡ các bác sĩ chẩn đoán bệnh nhanh chóng, chính xác hơn.

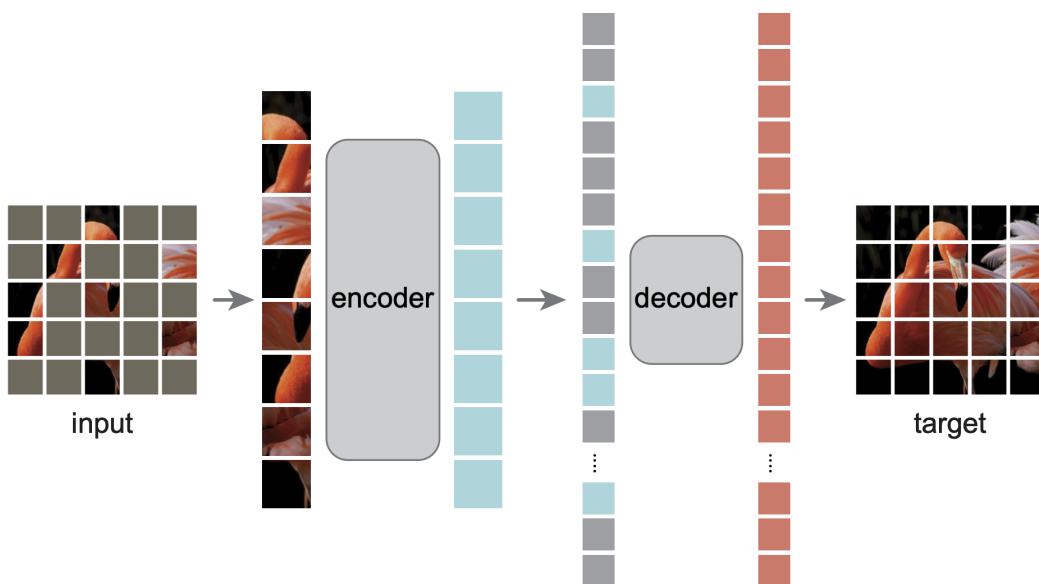
### 3.2 Tiền huấn luyện

Để có thể thực hiện được các nhiệm vụ mục tiêu, trước tiên cần phải đào tạo ra một xương sống mạnh mẽ bằng cách đào tạo bộ mã hóa theo phương pháp tự giám

sát trên lượng dữ liệu không nhãn với chiến lược che đi ngẫu nhiên các mảng của hình ảnh ban đầu. Khi một phần của hình ảnh bị che đi, mô hình tiền huấn luyện phải học cách dự đoán phần bị che dựa trên ngữ cảnh xung quanh. Điều này bắt buộc mô hình học các đặc trưng hữu ích và mạnh mẽ hơn từ dữ liệu. Việc mô hình phải dự đoán phần bị che giúp nó trở nên linh hoạt hơn trong việc nhận dạng và phân tích các đặc trưng của hình ảnh, từ đó cải thiện khả năng tổng quát hóa khi áp dụng vào các dữ liệu mới.

Trong bối cảnh như trên, chúng tôi sử dụng mô hình Masked Autoencoders nhằm tạo ra một mô hình nền tảng cho hình ảnh y tế. Mô hình Masked Autoencoders được giới thiệu lần đầu tại hội nghị CVPR2022 bởi He và cộng sự với bài báo có tên là "Masked Autoencoders are scalable vision learners" [4]. Mô hình đã chứng minh được tính hiệu quả trong việc học biểu diễn từ dữ liệu không có nhãn và khả năng học mạnh mẽ các đặc trưng của hình ảnh.

### Kiến trúc mô hình:



**Hình 3.2:** Kiến trúc mô hình Masked Autoencoders [4]

Mô hình có bộ mã hóa và giải mã đều là các mô hình Vision Transformer [9], trong đó số lớp trong bộ giải mã ít hơn so với số lớp trong bộ mã hóa. Để huấn luyện mô hình, các tác giả đã chỉ ra cách huấn luyện mô hình theo phương pháp học tự giám sát bằng cách chia hình ảnh thành các mảng và che đi ngẫu nhiên 75% các mảng của hình ảnh đầu vào rồi cho những mảng hình ảnh không bị che qua bộ mã hóa nhằm học cách "nén" những mảng không bị che thành các biểu diễn nhỏ hơn. Sau đó kết hợp những mảng bị che và biểu diễn đầu ra của bộ mã hóa theo thứ

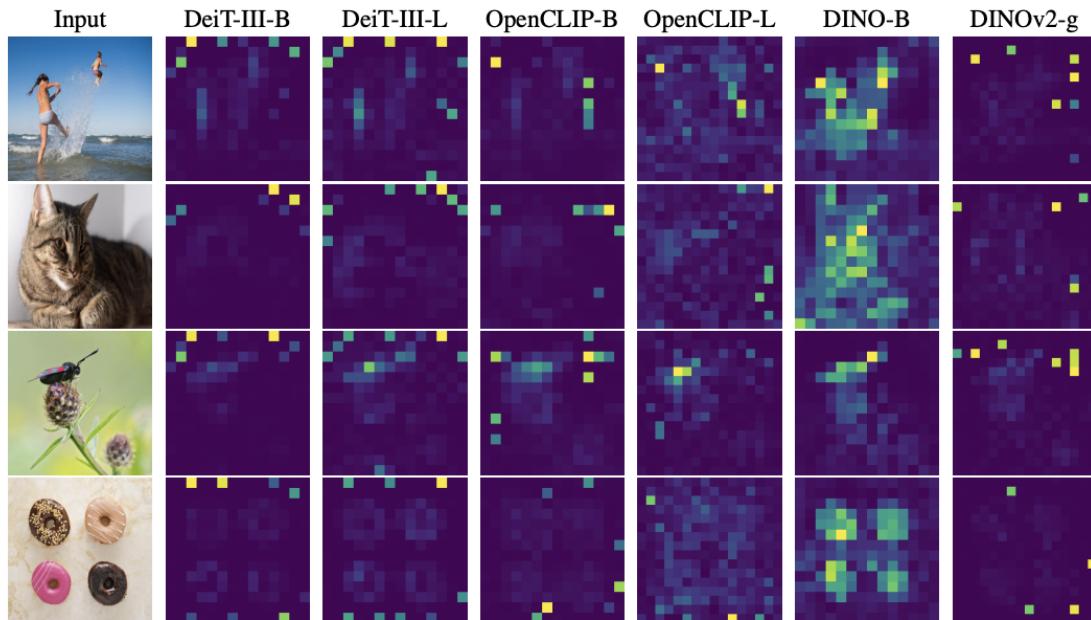
tự gốc ban đầu cho đi qua bộ giải mã để tái tạo lại hình ảnh gốc. Quá trình tự đào tạo từ dữ liệu không nhãn với nhiệm vụ giả định tái tạo hình ảnh theo pixel giúp mô hình có khả năng mạnh mẽ trong việc học hỏi các đặc trưng quan trọng của hình ảnh.

**Hàm mất mát:** Mô hình sử dụng hàm mất mát bình phương sai số trung bình (Mean Squared Error - MSE) giữa những mảng hình ảnh bị che ban đầu và những mảng hình ảnh được tái tạo từ bộ giải mã trong một lô (batch) dữ liệu để huấn luyện. Công thức của hàm mất mát như sau:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M \|\hat{X}_{ij} - X_{ij}\|_2^2 \quad (3.1)$$

Trong đó,  $\hat{X}_{ij}$  là mảng hình ảnh được tái tạo từ bộ giải mã,  $X_{ij}$  là mảng hình ảnh ban đầu,  $N$  là số lượng hình ảnh trong một batch,  $M$  là số lượng mảng hình ảnh trong một hình ảnh.

**ViT cần các token bổ sung:** Khi hình dung các hoạt động bên trong của Vision Transformers (ViT) [9], các nhà nghiên cứu nhận thấy sự chú ý tăng đột biến một cách kỳ lạ trên các mảng nền ngẫu nhiên. (xem Hình 3.3)



**Hình 3.3:** Hiển thị trực quan một số mảng nền gây ra sự chú ý trong DEiT [21], CLIP [22], DINOv2 [23] mà các nhà nghiên cứu muốn giảm bớt.

Nhiều nghiên cứu trước đây đã chỉ ra rằng các bản đồ chú ý do ViT [9] tạo ra muộn mà dễ hiểu. Việc hiển thị trực quan bản đồ chú ý cho phép chúng ta xem qua phần nào của hình ảnh mà mô hình đang tập trung vào. Tuy nhiên, nhiều biến

thể ViT lại thể hiện sự chú ý cao trên các mảng nền của hình ảnh nơi mà ở đó không có thông tin. Bằng cách trực quan hóa bản đồ chú ý trên các mô hình và tạo ra hình ảnh giống như Hình 3.3, các nhà nghiên cứu cho thấy rõ điều này xảy ra. Bằng cách thăm dò số lượng các phần nhúng đầu ra, các tác giả đã xác định được nguyên nhân gốc rễ. Một phần nhỏ (khoảng 2%) token mảng (patch token) có chuẩn hóa L2 (L2 Normalization) cao bất thường.

Việc chuẩn hóa L2 cao bất thường ở những token mảng có thể làm cho mô hình:

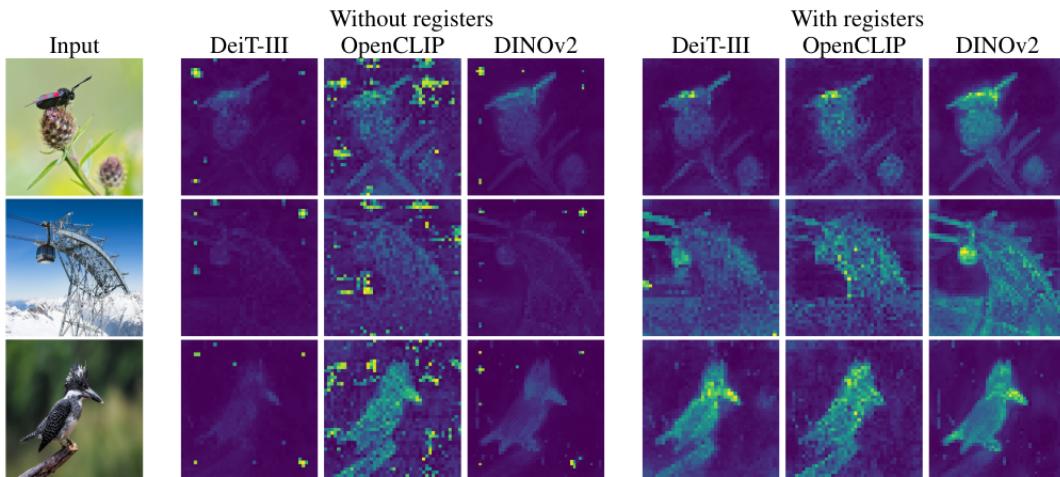
- Quá khớp (overfitting)
- Tính không ổn định về mặt số: Tham số mô hình rất lớn hoặc rất nhỏ có thể gây ra các vấn đề về số, dẫn đến không ổn định.
- Khả năng khái quát hóa kém: chuẩn hóa L2 cao cũng có thể chỉ ra rằng mô hình có thể không khái quát hóa tốt đối với dữ liệu mới, chưa được nhìn thấy.

Để có một cái nhìn cụ thể hơn, ta có một ví dụ như sau: giả sử ta đang cố gắng giữ thăng bằng cho một chiếc bập bênh và có các vật nặng với nhiều kích cỡ khác nhau để đặt ở hai bên của chiếc bập bênh. Kích thước hay khối lượng của mỗi vật thể hiện mức độ ảnh hưởng hoặc tầm quan trọng của nó trong việc cân bằng bập bênh. Nay giờ, nếu một trong những vật đó lớn bất thường, điều này có ý nghĩa rằng vật đó đang có quá nhiều ảnh hưởng đến sự cân bằng. Trong học sâu, điều này có thể làm lu mờ các phần quan trọng khác và có thể dẫn đến những quyết định sai lầm. Qua một vài thí nghiệm khác, các nhà nghiên cứu đã quan sát thấy những mảng hình ảnh ngoại lệ này xuất hiện trên các vùng rất giống với các vùng lân cận, điều này cho thấy sự dư thừa.

Các tác giả đã nhận ra giả thuyết rằng các mô hình học cách nhận ra các mảng hình ảnh chứa ít thông tin và mô hình đã "tái chế" các token tương ứng để tổng hợp thông tin hình ảnh toàn cầu trong khi loại bỏ thông tin không gian. Điều này cho phép xử lý đặc trưng nội bộ hiệu quả. Tuy nhiên, việc tái chế này gây ra những tác dụng phụ không mong muốn:

- Mất đi chi tiết của mảng hình ảnh ban đầu và làm ảnh hưởng đến các nhiệm vụ như phân đoạn
- Bản đồ chú ý khó diễn giải

Để giảm bớt các mảng hình ảnh được tái chế, các nhà nghiên cứu đề xuất cung cấp cho các mô hình một bộ nhớ chuyên dụng bằng cách thêm các token bổ sung (register token) bên cạnh token cls. Điều này cung cấp không gian tạm thời cho các tính toán nội bộ, ngăn chặn việc chiếm đoạt các phần nhúng của các mảng ngẫu nhiên.

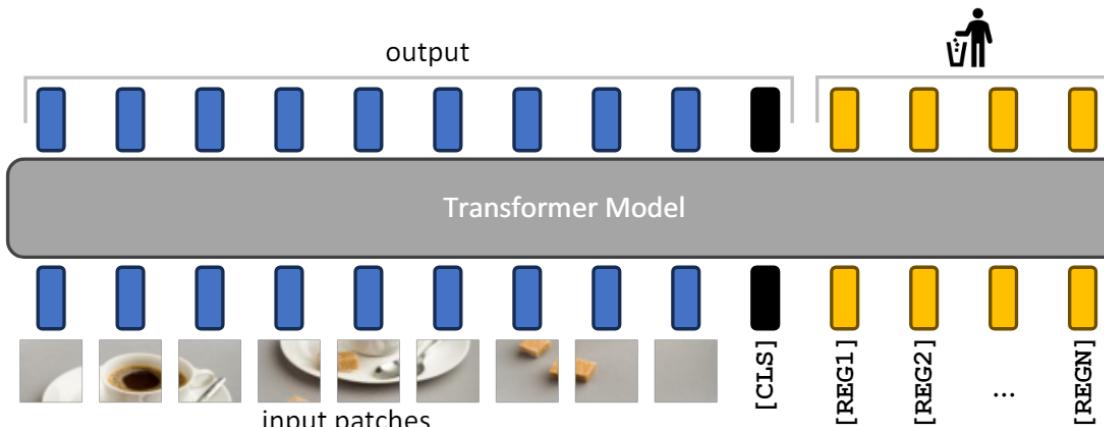


**Hình 3.4:** Ảnh hưởng của các token bổ sung (register token) giúp ViT chú ý chính xác đến đối tượng

Điều đáng chú ý là tinh chỉnh đơn giản này hoạt động rất tốt. Các mô hình được đào tạo với các token bổ sung này (register token) đem lại:

- Bản đồ chú ý mượt mà và có ý nghĩa hơn về mặt ngữ nghĩa.
- Tăng hiệu suất nhỏ trên các điểm chuẩn khác nhau.

Chỉ cần một thay đổi nhỏ về kiến trúc sẽ mang lại những lợi ích đáng chú ý (xem Hình 3.4). Vì vậy, với những phát hiện trên của nhóm tác giả nghiên cứu bài báo "Vision Transformers Need Registers" [24], chúng tôi đã thử nghiệm việc thêm vào trong mô hình MAE [4] các token bổ sung trong giai đoạn tiền huấn luyện. Khi huấn luyện mô hình trên các nhiệm vụ mục tiêu, cũng giống như token cls, các token bổ sung (register token) sẽ bị bỏ đi khi chỉ lấy đầu ra của các mảng ảnh ban đầu (xem Hình 3.5).



**Hình 3.5:** Minh họa cho việc thêm register token vào MAE [4]

### 3.3 Tinh chỉnh trên các nhiệm vụ mục tiêu

Mô hình Masked Autoencoders [4] sau khi được tiền huấn luyện sẽ được tinh chỉnh trên các nhiệm vụ mục tiêu như phân vùng và phân loại. Vì mô hình đã học được cách biểu diễn mạnh mẽ dữ liệu bằng cách tự đào tạo sử dụng phương pháp tự giám sát che giấu ngẫu nhiên 75% các mảng của hình ảnh nên ở giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu, ta sẽ bỏ đi bộ giải mã của mô hình và thay thế bằng một đầu ra mới phù hợp với nhiệm vụ cụ thể.

#### 3.3.1 Mô hình phân vùng hình ảnh nội soi y tế

Sau quá trình tiền huấn luyện thu được một bộ mã hóa (một mô hình Vision Transformers [9]) từ mô hình Masked Autoencoders [4], cần có một bộ giải mã phù hợp cho từng nhiệm vụ mục tiêu cụ thể. Ở đây chúng tôi xem xét nhiệm vụ mục tiêu là phân vùng các đối tượng trong ảnh.

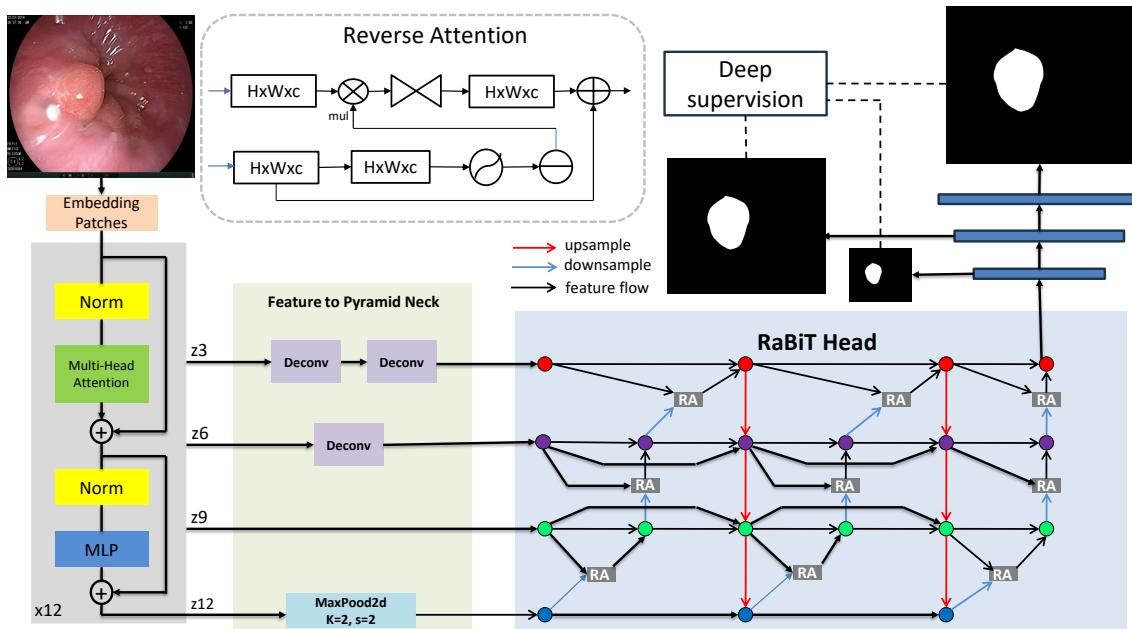
##### Bộ giải mã RaBiT [25]:

Sau khi có được bộ mã hóa đã được tiền huấn luyện với hình ảnh nội soi, chúng tôi sử dụng một bộ giải mã mạnh mẽ của mô hình RaBiT [20] để tạo thành mô hình dạng bộ mã hóa - giải mã (encoder - decoder) cho nhiệm vụ phân vùng tổn thương trong hình ảnh nội soi y tế.

Bộ giải mã RaBiT được thiết kế để tận dụng điểm mạnh của cả mô-đun BiFPN (Bidirectional Feature Pyramid Network) [25] và RA (Reverse Attention) [18]. Ban đầu, RA được sử dụng theo hướng từ dưới lên để tinh chỉnh ranh giới polyp giúp cải thiện độ chính xác của phân vùng. RaBiT đã sử dụng RA để tinh chỉnh lặp đi lặp lại các ranh giới polyp. Bộ giải mã kết hợp giữa tổng hợp các đặc trưng từ trên xuống và sàng lọc ranh giới polyp từ dưới lên dưới dạng mô-đun Mạng Kim tự tháp Tính năng hai chiều dựa trên Chú ý ngược (RaBiFPN), được lặp lại nhiều lần để tạo thành một khối. Trong đó, mô-đun RaBiFPN được lặp lại 4 lần. Tất cả các bản đồ đặc trưng trong bộ giải mã được tổng hợp bằng cách hợp nhất bình thường hóa nhanh [25], trong đó trọng số hợp nhất được học bằng cách lan truyền ngược trong giai đoạn huấn luyện. Cuối cùng, các tác giả kết hợp một nhánh tinh chỉnh cuối cùng để tạo ra dự đoán cuối cùng. RaBiT cải tiến RA để xử lý bản đồ đặc trưng nhiều kênh  $H \times W \times 224$ . Trong phân đoạn nhị phân, một lớp tích chập 3x3 nén bản đồ đặc trưng thành một kênh và áp dụng hàm sigmoid để tạo ra bản đồ chú ý ngược. Từ các cải tiến của RaBiT, các kết quả thử nghiệm của bài báo này đã cho một hiệu suất cao trong phân vùng ảnh, chúng tôi quyết định sử dụng bộ giải mã của RaBit và so sánh hiệu suất giữa mô hình RaBiT và mô hình của chúng tôi trên các điểm chuẩn polyp ở Chương 4.5.

**Mô-đun Feature to Pyramid:** Trong bộ mã hóa Vision Transformer, đặc trưng ở các lớp có kích thước giống nhau. Điều này làm cho các đặc trưng này không thể truyền thẳng vào bộ giải mã của RaBiT. Vì thế, chúng tôi đã sử dụng một mô-đun giúp thay đổi từ kích thước đặc trưng giống nhau ở từng lớp thành các kích thước khác nhau với phép biến đổi lấy mẫu lên (upsample) và lấy mẫu xuống (downsample). Đó là mô-đun chuyển đổi đặc trưng thành dạng kim tự tháp F2P (Feature to Pyramid) (xem Hình 3.6).

### Kiến trúc mô hình:



Hình 3.6: Kiến trúc mô hình phân vùng hình ảnh nội soi y tế

Cuối cùng, sau khi có được mô-đun chuyển đổi đặc trưng thành dạng kim tự tháp, bộ mã hóa và bộ giải mã, chúng tôi kết hợp chúng lại thành một mô hình dùng cho nhiệm vụ phân đoạn hình ảnh (xem Hình 3.6). Ở trong giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu, hình ảnh đầu vào sẽ không bị che đi. Mô hình được huấn luyện trên nhiều độ phân giải (multi-scale) khác nhau của hình ảnh đầu vào, từ đó giúp cho mô hình có khả năng học tổng quát. Bên cạnh đó, mô hình cũng kết hợp giám sát sâu (deep supervision [26]) ở nhiều độ phân giải của kết quả đầu ra của mô hình và tính toán mất mát trên nhiều độ phân giải đó nhằm giúp mô hình học được đa dạng các đặc trưng của hình ảnh khi ở độ phân giải cao thì mức độ chi tiết của hình ảnh cũng càng cao.

### Hàm mất mát:

Hàm mất mát được sử dụng để huấn luyện mô hình là sự kết hợp từ hai hàm mất mát có trọng số Focal và hàm mất mát có trọng số IoU. Hàm mất mát Focal tập trung vào phân phối pixel trên bản đồ đầu ra, chú trọng từng pixel riêng lẻ. Do

vấn đề trong phân vùng hình ảnh y tế cho các loại bệnh thường không phải lúc nào chũng chiếm đúng bằng một nửa hình ảnh mà có những loại bệnh như polyp thì vùng tổn thương chỉ chiếm một phần nhỏ trong hình ảnh đầu vào, trong khi phần lớn còn lại là nền, dẫn đến bản đồ ban đầu bị mất cân bằng. Ngược lại, hàm mất mát weighted IoU là một hàm mất mát tập trung vào từng vùng pixel để nắm bắt mối tương quan giữa các pixel lân cận.

Nhóm tác giả đã giả thiết rằng có một số pixel quan trọng hơn những pixel khác khi đóng góp vào quá trình huấn luyện. Trọng số  $\beta_{ij}$  cho pixel  $(i, j)$  đại biểu cho mối quan hệ của pixel trung tâm và lân cận của nó:

$$\beta_{ij} = \left| \frac{\sum_{m,n \in N_{ij}} g_{mn}}{|N_{ij}|} - g_{ij} \right|$$

Trong đó  $N_{ij}$  là vùng lân cận  $31 \times 31$  của pixel  $(i, j)$  và  $g_{ij}$  là nhãn của pixel  $(i, j)$ . Phương pháp đánh trọng số như vậy sẽ cho phép mô hình tập trung nhiều hơn vào vùng biên.

Giả sử  $p_{i,j}$  là dự đoán xác suất của mô hình cho pixel  $(i, j)$ , ta có  $q_{i,j}$ :

$$q_{i,j} = \begin{cases} p_{i,j}, & \text{if } g_{ij} = 1 \\ 1 - p_{i,j}, & \text{otherwise} \end{cases}$$

Như vậy công thức weighted focal loss và weighted IoU loss sẽ được biểu diễn như sau:

$$L_{w\_focal} = -\frac{\sum_{i=1}^H \sum_{j=1}^W (1 + \lambda \beta_{ij}) \alpha (1 - q_{ij})}{\sum_{i=1}^H \sum_{j=1}^W (1 + \lambda \beta_{ij})}$$

$$L_{w\_iou} = 1 - \frac{\sum_{i=1}^H \sum_{j=1}^W (g_{ij} * p_{ij}) * (1 + \lambda \beta_{ij})}{\sum_{i=1}^H \sum_{j=1}^W (g_{ij} + p_{ij} * p_{ij}) * (1 + \lambda \beta_{ij})}$$

Trong đó  $\gamma$ ,  $\alpha$  là các tham số có thể tinh chỉnh.

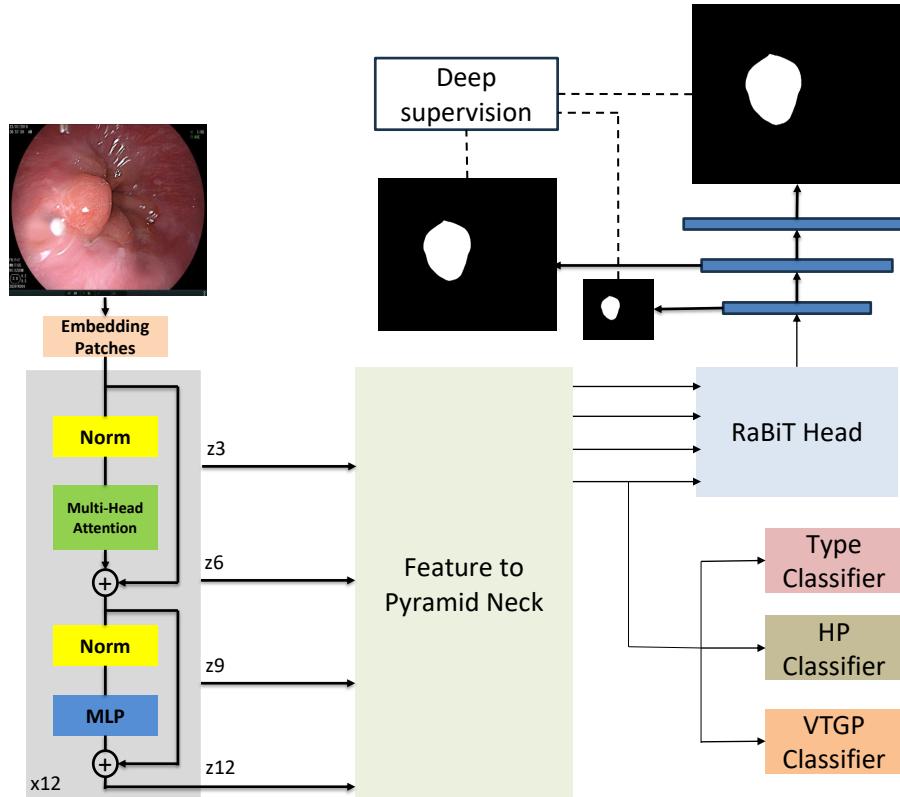
Cuối cùng, hàm loss sẽ được tính dựa trên trung bình cộng của 2 hàm loss kể trên:

$$L_{total} = \frac{L_{w\_focal} + L_{w\_iou}}{2}$$

Mô hình cho ra nhiều bản đồ dự đoán với nhiều multi scale, do đó hàm loss

$L_{total}$  ở công thức 3.3.1 sẽ được áp dụng cho từng mức scale và cộng tổng lại để tính gradient.

### 3.3.2 Mô hình đa nhiệm



Hình 3.7: Kiến trúc mô hình đa nhiệm

Ngoài mô hình đơn nhiệm, chúng tôi đề xuất một mạng đa nhiệm cho ảnh nội soi y tế. Mô hình đa nhiệm có dạng kiến trúc giống mô hình EndoUNet bao gồm một đầu phân đoạn hình ảnh y tế giống như ở Chương 3.3.1 kết hợp thêm ba đầu phân loại HP lành tính hay ác tính, phân loại các vị trí giải phẫu và phân loại bệnh bằng các mạng nơ-ron tuyển tính (xem Hình 3.7). Mô hình được huấn luyện trên một độ phân giải duy nhất 384x384 và được tính toán mắt mát với giám sát sâu (deep supervision). Hàm mắt mát được sử dụng để huấn luyện mô hình là tổng của hàm mắt mát phân đoạn (xem 3.3.1), một hàm phân loại nhị phân cross entropy và hai hàm mắt mát phân loại cross entropy cho hai nhiệm vụ tương ứng là phân loại loại bệnh và phân loại vị trí giải phẫu.

Phân loại HP lành tính hay ác tính sử dụng hàm mắt mát nhị phân cross entropy như sau:

$$\mathcal{L}_{\text{binary}} = -\frac{1}{N} \sum_{i=1}^N y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \quad (3.2)$$

Phân loại vị trí giải phẫu và phân loại loại bệnh sử dụng hàm mắt mát cross

entropy như sau:

$$\mathcal{L}_{\text{cross}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}) \quad (3.3)$$

Hàm mất mát tổng cộng của mô hình đa nhiệm được tính như sau:

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{hp}} + \mathcal{L}_{\text{vitrigiaiphau}} + \mathcal{L}_{\text{loaibenh}} \quad (3.4)$$

## CHƯƠNG 4. ĐÁNH GIÁ THỰC NGHIỆM

Chương này mô tả các bộ dữ liệu cho các nhiệm vụ tiền huấn luyện, phân vùng và phân loại tổn thương. Cung cấp thông tin chi tiết về từng tập dữ liệu, chuẩn bị dữ liệu và chiến lược tăng cường dữ liệu cho quá trình tiền huấn luyện và các nhiệm vụ mục tiêu. Tất cả dữ liệu được sử dụng trong đồ án này (ngoại trừ các tập dữ liệu công khai) đều là dữ liệu thực tế được thu thập từ kết quả nội soi của bệnh nhân tại bệnh viện ở Hà Nội. Trong chương này, cũng cung cấp thông tin chi tiết về cài đặt siêu tham số cho các thử nghiệm. Cuối cùng, giới thiệu về các độ đo để đánh giá và thảo luận về các kết quả thử nghiệm thu được.

### 4.1 Dữ liệu

**Dữ liệu được trích xuất từ video** là những hình ảnh được dự đoán có tổn thương hoặc dương tính giả (đối với polyp) với một ngưỡng phát hiện nhỏ hơn ngưỡng dự đoán có tổn thương từ mô hình phát hiện tổn thương tiêu hóa. Chúng tôi đã sử dụng các mô hình phát hiện tổn thương tiêu hóa đã được huấn luyện trên các tập dữ liệu polyp và tổn thương tiêu hóa như YOLOv8 [27] và mô hình MAE-UNETR [28] cũ của chúng tôi nhằm tạo thêm lượng dữ liệu không nhãn cho quá trình tiền huấn luyện. Khi trích xuất ảnh từ video, chúng tôi đã chọn khoảng cách 5 khung hình lấy 1 khung hình để dự đoán. Sau khi trích xuất dữ liệu từ video, trong những hình ảnh thu được, có một vài hình ảnh có nhiều nước, mờ, ... Chúng tôi đã lọc và loại bỏ thủ công những hình ảnh này để tạo ra một tập dữ liệu chất lượng. Bảng 4.1 thống kê chi tiết về dữ liệu được trích xuất từ video.

**Bảng 4.1:** Dữ liệu được trích xuất từ video

Loại dữ liệu	Số lượng
Dữ liệu polyp	8981
Dữ liệu polyp dương tính giả	15303
Dữ liệu tổn thương tiêu hóa	9113
Tổng cộng	33397

**Dữ liệu polyp công khai** bao gồm hai tập dữ liệu huấn luyện và kiểm thử của mô hình PraNet [17]. Bảng 4.2 và Bảng 4.3 thống kê chi tiết về dữ liệu.

**Bảng 4.2:** Dữ liệu huấn luyện mô hình PraNet [17]

Loại dữ liệu	Số lượng
Kvasir-SEG	900
CVC-ClinicDB	550
Tổng cộng	1450

**Bảng 4.3:** Dữ liệu kiểm thử mô hình PraNet [17]

Loại dữ liệu	Số lượng
CVC-300	60
CVC-ClinicDB	62
CVC-ColonDB	380
ETIS-Larib Polyp DB	196
Kvasir-SEG	100

**Dữ liệu polyp từ bệnh viện** là những hình ảnh nội soi về polyp được thu thập và được gán nhãn bởi các bác sĩ chuyên khoa. Dữ liệu được chia thành hai tập: tập huấn luyện và tập kiểm thử. Bảng 4.4 thống kê chi tiết về dữ liệu.

**Bảng 4.4:** Dữ liệu polyp từ bệnh viện

Loại dữ liệu	Số lượng
Tập huấn luyện	8561
Tập kiểm thử	2000

**Dữ liệu tổn thương tiêu hóa trên** được thu thập từ bệnh nhân trong bệnh viện. Dữ liệu đã được các bác sĩ chuyên khoa gán nhãn phân vùng cho mỗi tổn thương tiêu hóa. Dữ liệu được chia thành hai phần: dữ liệu huấn luyện và dữ liệu kiểm thử được chia theo tỷ lệ 80% và 20%. Bảng 4.5 thống kê tất cả các dữ liệu về các tổn thương của đường tiêu hóa.

**Bảng 4.5:** Dữ liệu tiền huấn luyện

Loại dữ liệu	Số lượng huấn luyện	Số lượng kiểm thử
Ung thư dạ dày	1481	371
Ung thư thực quản	984	247
Viêm dạ dày	4189	1047
Viêm thực quản	1989	498
Viêm loét hành tá tràng	944	237
Tổng cộng	9587	2400

**Dữ liệu phân loại HP** được sử dụng cho nhiệm vụ phân loại trong mô hình đa nhiệm ở Chương 3.3.2. Dữ liệu phân loại HP cũng được chia thành hai phần: dữ liệu huấn luyện và dữ liệu kiểm thử. Hai phần của dữ liệu này có số lượng bằng với số lượng hai phần dữ liệu về viêm dạ dày.

**Dữ liệu phân loại vị trí giải phẫu** được sử dụng cho nhiệm vụ phân loại trong mô hình đa tác vụ được đề cập ở Chương 3.3.2. Dữ liệu phân loại vị trí giải phẫu bao gồm 10 vị trí giải phẫu khác nhau. Dữ liệu được chia thành hai phần: dữ liệu huấn luyện và dữ liệu kiểm thử. Bảng 4.6 thống kê chi tiết về dữ liệu.

**Bảng 4.6:** Dữ liệu huấn luyện và kiểm thử phân loại vị trí giải phẫu

Vị trí	Số lượng huấn luyện	Số lượng kiểm thử
Phình vị	443	111
Hành tá tràng	448	112
Tam vị	444	111
Bờ cong lớn	442	111
Hang vị	443	111
Bờ cong nhỏ	444	111
Tá tràng	447	112
Thực quản	442	111
Thân vị	442	111
Hầu họng	440	110
Tổng cộng	4435	1111

**Dữ liệu tiền huấn luyện** bao gồm tất cả dữ liệu trích xuất từ video, dữ liệu huấn luyện của tập dữ liệu polyp công khai, dữ liệu huấn luyện polyp từ bệnh viện, dữ liệu huấn luyện phân đoạn tổn thương tiêu hóa trên, dữ liệu huấn luyện phân loại vị trí giải phẫu và một số hình ảnh nội soi khác không có nhãn của bệnh viện. Bảng 4.7 thông kê chi tiết về dữ liệu.

**Bảng 4.7:** Dữ liệu tiền huấn luyện

Loại dữ liệu	Số lượng
Dữ liệu trích xuất từ video	33397
Dữ liệu huấn luyện polyp công khai	1450
Dữ liệu huấn luyện polyp từ bệnh viện	8561
Dữ liệu huấn luyện phân đoạn tổn thương tiêu hóa	9587
Dữ liệu huấn luyện phân loại vị trí giải phẫu	4435
Dữ liệu hình ảnh nội soi khác	9390
Tổng	66820

## 4.2 Tiền xử lý dữ liệu và tăng cường dữ liệu

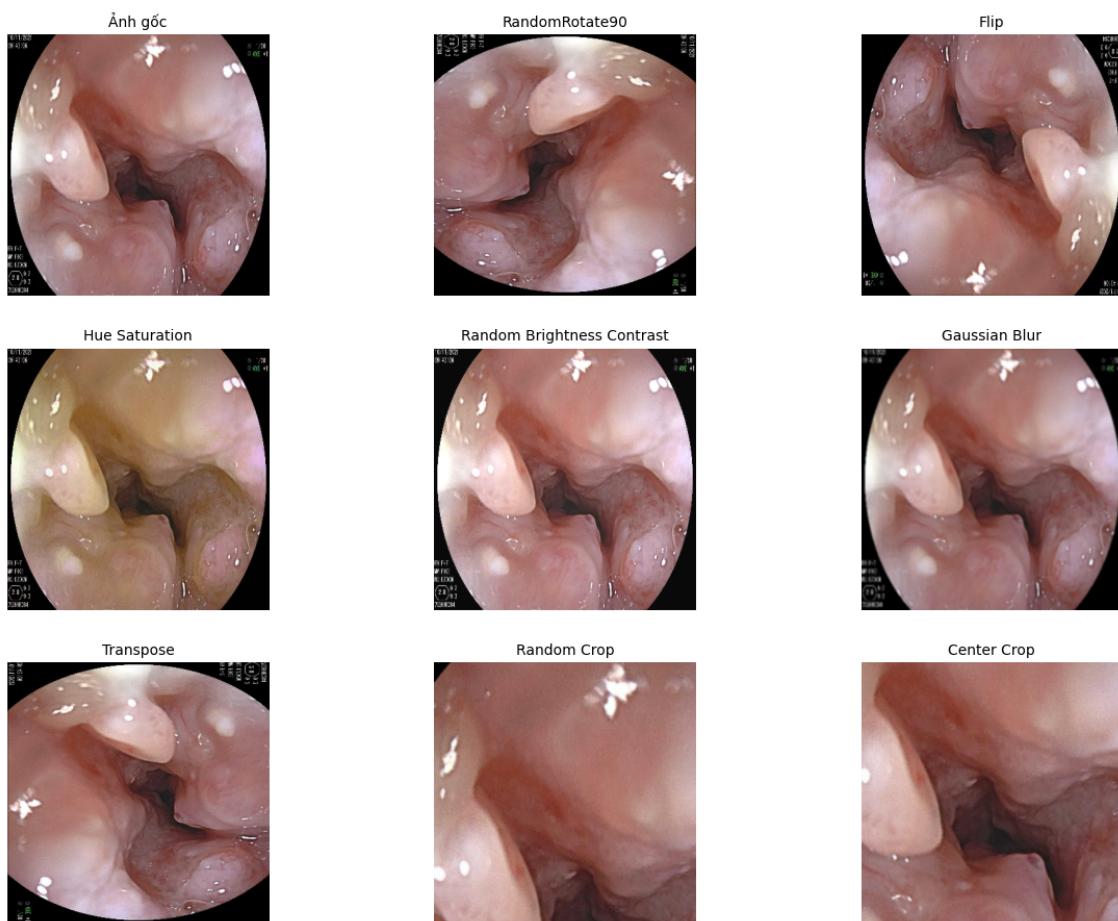
Vì các hình ảnh có kích thước khác nhau nên hình ảnh cần được thay đổi kích thước sao cho phù hợp trước khi đưa vào huấn luyện trong từng giai đoạn tiền huấn luyện hoặc giai đoạn huấn luyện các nhiệm vụ phía sau.

**Tiền huấn luyện:** Hình ảnh được cắt ngẫu nhiên theo kích thước về 224x224 trước khi đưa vào mô hình. Kỹ thuật tăng cường dữ liệu được sử dụng trong giai đoạn tiền huấn luyện chỉ có phép lật ngang ngẫu nhiên với xác suất 50%.

**Tinh chỉnh trên các nhiệm vụ mục tiêu:** Trong các nhiệm vụ mục tiêu, ngoài mô hình đơn nhiệm phân vùng ảnh, các hình ảnh sẽ được thay đổi thành nhiều kích thước khác nhau (256, 384, 512) thì các hình ảnh đều được thay đổi kích thước về

một kích thước duy nhất là  $384 \times 384$ . Việc tăng kích thước hình ảnh trong giai đoạn tinh chỉnh so với kích thước hình ảnh trong giai đoạn tiền huấn luyện ( $224 \times 224$ ) trên các nhiệm vụ mục tiêu giúp mô hình có thể cải thiện hiệu suất tốt hơn [29]. Dữ liệu huấn luyện được tăng cường theo thời gian thực với xác suất 0,5 (tức là mỗi hình ảnh có 50% cơ hội được tăng cường khi được chọn để huấn luyện). Các kỹ thuật sau đây được sử dụng (xem Hình 4.1):

- Phép xoay ngẫu nhiên 90 độ với số lần xoay lớn hơn hoặc bằng 0.
- Phép đảo ảnh (đảo ngang hoặc đảo dọc hoặc cả hai)
- Thay đổi ngẫu nhiên màu sắc, độ bão hòa và giá trị của hình ảnh đầu vào.
- Thay đổi ngẫu nhiên độ sáng và độ tương phản của hình ảnh đầu vào.
- Làm mờ hình ảnh đầu vào bằng nhiễu Gaussian.
- Lật hình ảnh qua đường chéo
- Cắt ngẫu nhiên ảnh với kích thước  $224 \times 224$  với xác suất cắt là 20%.
- Cắt chính giữa ảnh với kích thước  $224 \times 224$  và xác suất cắt là 20%.



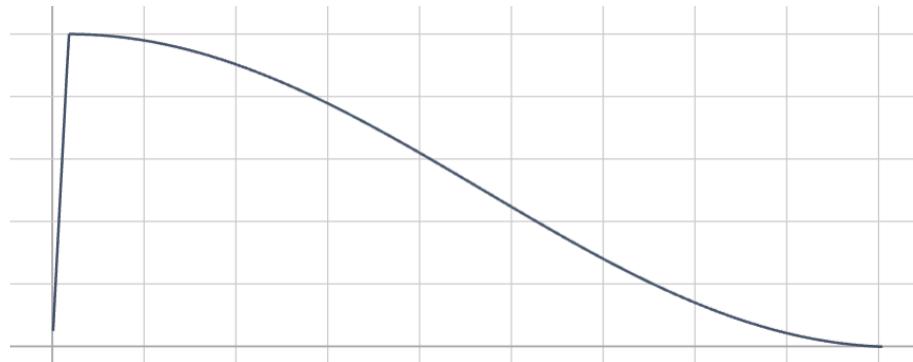
**Hình 4.1:** Một số kỹ thuật tăng cường dữ liệu được sử dụng

### 4.3 Triển khai chi tiết

**Môi trường triển khai:** Thí nghiệm được thực hiện trên máy chủ có cấu hình và môi trường như sau:

- Hệ điều hành: Ubuntu 20.04 LTS
- CPU: Intel Xeon E5-2620 v4 @ 2.10GHz
- RAM: 64GB
- GPU: NVIDIA RTX 3090 24GB
- CUDA: 12.1
- CUDNN: 8.0.5
- Python: 3.10
- PyTorch: 2.2.0
- Logs: TensorBoard

**Trình lập lịch Cosine:** Trong các giai đoạn huấn luyện, các mô hình đều sử dụng trình lập lịch khởi động tốc độ học Cosine như **Hình 4.2**.



**Hình 4.2:** Lập lịch khởi động tốc độ học Cosine

**Cài đặt siêu tham số:** Các siêu tham số trong giai đoạn tiền huấn luyện và giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu được cài đặt như Bảng 4.8 và Bảng 4.9.

### 4.4 Các độ đo

Trong phần này, chúng ta sẽ tìm hiểu về các độ đo được sử dụng để đánh giá trong đồ án này. Cụ thể, chúng ta sẽ tìm hiểu về độ đo Dice, IoU và Accuracy. Trước khi đi vào chi tiết, chúng ta có các kí hiệu như sau:

- A và B là hai tập hợp cần đo lường độ tương đồng.
- $A_i$  và  $B_i$  là hai tập hợp cần đo lường độ tương đồng của lớp  $i$ .
- $|A|$  và  $|B|$  là số phần tử của tập hợp A và B.

**Bảng 4.8:** Cài đặt siêu tham số cho tiếp tục tiền huấn luyện.

Kích thước lô	256
Tích lũy gradient	1
Tốc độ học	1e-4
Kỉ nguyên khởi động	1
Kỉ nguyên huấn luyện	100
Trình tối ưu hóa	AdamW
Betas	(0.9, 0.95)
Chuẩn hóa theo pixel	Có

**Bảng 4.9:** Cài đặt siêu tham số cho các nhiệm vụ mục tiêu.

Kích thước lô	64
Tích lũy gradient	1
Tốc độ học	1e-4
Kỉ nguyên khởi động	1
Kỉ nguyên huấn luyện	20
Trình tối ưu hóa	AdamW
Betas	(0.9, 0.99)

- $|A \cap B|$  là số phần tử chung của tập hợp A và B.
- $|A \cup B|$  là tổng số phần tử của tập hợp A và B.
- TP (True Positive) là số lượng các trường hợp mà mô hình dự đoán đúng.
- TN (True Negative) là số lượng các trường hợp mà mô hình dự đoán sai.
- FP (False Positive) là số lượng các trường hợp mà mô hình dự đoán đúng nhưng thực tế là sai.
- FN (False Negative) là số lượng các trường hợp mà mô hình dự đoán sai nhưng thực tế là đúng.
- Macro: Tính trung bình của các giá trị đo lường trên từng lớp.
- Micro: Tính trung bình của các giá trị đo lường trên toàn bộ dữ liệu.

#### 4.4.1 Độ đo Dice

Độ đo Dice đo lường sự giống nhau giữa hai tập hợp. Trong bài toán phân đoạn ảnh y khoa, độ đo Dice được sử dụng để đo lường sự giống nhau giữa ảnh gốc và ảnh được dự đoán từ mô hình.

Độ đo Dice được định nghĩa như sau:

$$DSC = \frac{2 \times |A \cap B|}{|A| + |B|} \quad (4.1)$$

**Macro Dice:** Đối với mỗi lớp  $i$ , độ đo Dice ( $DSC_i$ ) được tính như sau:

$$DSC_i = \frac{2 \times |A_i \cap B_i|}{|A_i| + |B_i|} \quad (4.2)$$

Chỉ số Macro Dice được tính bằng cách lấy trung bình của các chỉ số Dice của tất cả các lớp:

$$\textbf{Macro Dice} = \frac{1}{n} \sum_{i=1}^n DSC_i \quad (4.3)$$

**Micro Dice:** Micro Dice là một biến thể của chỉ số Dice được sử dụng để đánh giá hiệu suất của các mô hình phân đoạn. Khác với Macro Dice, Micro Dice tính toán mức độ tương đồng dựa trên tổng thể các pixel của tất cả các lớp, thay vì trung bình chỉ số Dice của từng lớp. Tổng số pixel trong vùng giao nhau, tổng số pixel thực tế và tổng số pixel dự đoán của tất cả các lớp được tính như sau:

$$\begin{aligned} \textbf{MI} &= \sum_{i=1}^n |A_i \cap B_i| \\ \textbf{MA} &= \sum_{i=1}^n |A_i| \\ \textbf{MP} &= \sum_{i=1}^n |B_i| \end{aligned} \quad (4.4)$$

Chỉ số Micro Dice được tính bằng công thức:

$$\textbf{Micro Dice} = \frac{2 \times MI}{MA + MP} \quad (4.5)$$

#### 4.4.2 Độ đo IoU

Cũng giống như độ đo Dice, độ đo IoU (hay độ đo Jaccard) cũng được sử dụng để đo lường sự giống nhau giữa hai tập hợp. Độ đo Dice nhấn mạnh sự trùng lặp bằng cách nhân đôi nó trong tử số, trong khi độ đo IoU chỉ sử dụng giao chia cho hợp. Độ đo IoU được định nghĩa như sau:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (4.6)$$

**Macro IoU:** Độ đo IoU của lớp  $i$  được tính như sau:

$$IoU_i = \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (4.7)$$

Độ đo Macro IoU được tính bằng cách lấy trung bình của các độ đo IoU của tất cả

các lớp:

$$\text{Macro IoU} = \frac{1}{n} \sum_{i=1}^n IoU_i \quad (4.8)$$

**Micro IoU:** Tổng số pixel trong vùng giao nhau, tổng số pixel thực tế và pixel dự đoán của tất cả các lớp được tính như sau:

$$\begin{aligned} \mathbf{MI} &= \sum_{i=1}^n |A_i \cap B_i| \\ \mathbf{MU} &= \sum_{i=1}^n |A_i \cup B_i| \end{aligned} \quad (4.9)$$

Chỉ số Micro IoU được tính bằng công thức:

$$\text{Micro IoU} = \frac{\mathbf{MI}}{\mathbf{MU}} \quad (4.10)$$

#### 4.4.3 Độ chính xác

Độ chính xác là một độ đo thống kê được sử dụng để đánh giá hiệu suất của một mô hình phân loại. Độ chính xác được tính bằng tỉ lệ giữa số lượng dự đoán đúng và tổng số lượng dự đoán. Công thức tính độ chính xác như sau:

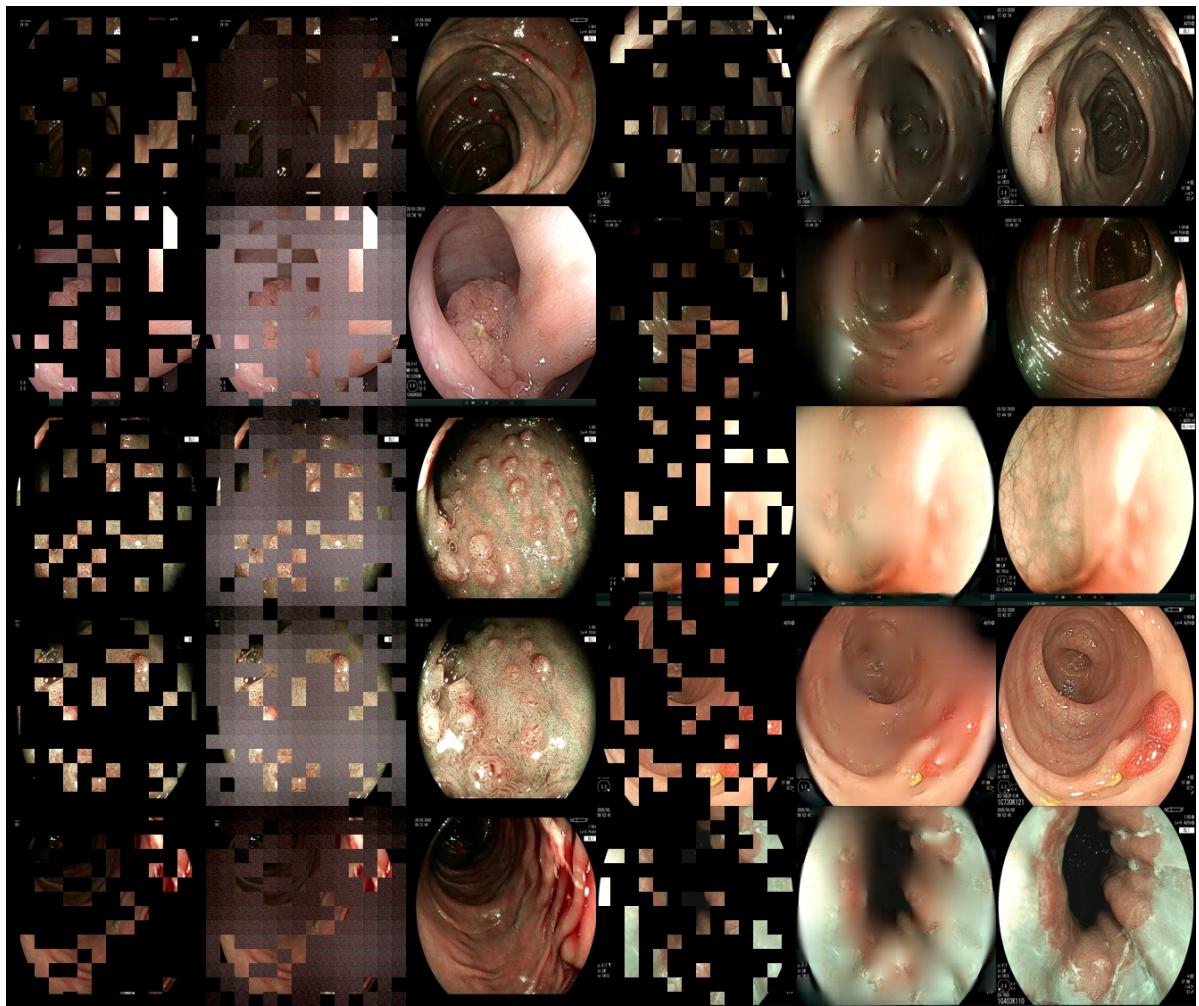
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.11)$$

### 4.5 Kết quả thử nghiệm

Trong phần này, sau quá trình tiền huấn luyện, chúng tôi thử nghiệm tinh chỉnh mô hình và so sánh hiệu suất của mô hình trên các nhiệm vụ mục tiêu với các mô hình được đề xuất ở Chương 3.3 và dữ liệu được đề cập ở Chương 4.1.

#### 4.5.1 Chất lượng tái tạo hình ảnh

Hình 4.3 thể hiện chất lượng tái tạo sử dụng phương pháp trong mô hình Masked Autoencoders [4] trong kỷ nguyên đầu và kỷ nguyên cuối cùng của quá trình huấn luyện. Nhìn chung, ta có thể quan sát thấy mô hình có thể tái tạo lại hình ảnh bị che ngẫu nhiên 75% với chất lượng tương đối tốt. Điều này cho thấy bộ mã hóa sau khi đào tạo trước, có thể trích xuất các đặc trưng có ý nghĩa từ hình ảnh. Cũng có thể nhận thấy rằng chất lượng tái tạo hình ảnh về mặt khách quan không thể giống như ảnh gốc ban đầu, vì vậy, ở giai đoạn tinh chỉnh cho các nhiệm vụ phía sau, tôi quyết định không đóng băng bộ mã hóa mà để mô hình học được các đặc trưng mới từ dữ liệu khi hình ảnh đầu vào không còn bị che.



**Hình 4.3:** Minh họa chất lượng tái tạo hình ảnh sử dụng phương pháp của Masked Autoencoders [4] ở epoch 1 và ở epoch cuối cùng. Cột đầu tiên là những hình ảnh bị che đi 75%, cột thứ hai là những hình ảnh mà mô hình tái tạo lại và cột thứ ba những hình ảnh ban đầu trước khi bị che

Tuy nhiên, theo tác giả của bài báo Masked Autoencoders [4] đã nêu ra<sup>1</sup>, trong giai đoạn tiền huấn luyện, nếu hình ảnh được mô hình tái tạo không chuẩn hóa theo pixel trước khi tính toán với hàm mất mát thì mô hình cho hiệu suất ở giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu thấp hơn so với mô hình tính toán mất mát trên hình ảnh tái tạo được chuẩn hóa theo pixel.

#### 4.5.2 Phân đoạn hình ảnh nội soi

Trong phần này, chúng tôi sẽ thử nghiệm trên mô hình cơ sở của MAE (mô hình đã được tiền huấn luyện trên ImageNet-1k [7]), mô hình MAE được tiền huấn luyện từ đầu, mô hình cơ sở MAE được **tiếp tục** tiền huấn luyện và mô hình cơ sở MAE có thêm bốn token bổ sung (register token![24]) được tiếp tục tiền huấn luyện và so sánh với các mô hình có giám sát từ đầu khác. Bảng 4.10, Bảng 4.11 và Bảng 4.12 là kết quả mà chúng tôi thu được sau khi thử nghiệm với **MAE-Scratch-**

<sup>1</sup><https://github.com/facebookresearch/mae/issues/12#issuecomment-1011689674>

**RaBiFPN** là sự kết hợp giữa bộ giải mã MAE được tiền huấn luyện từ đầu và bộ giải mã RaBiT (mô-đun RaBiFPN); **MAE-Base-RaBiFPN** là sự kết hợp giữa bộ giải mã MAE được tiền huấn luyện trên ImageNet-1k và bộ giải mã RaBiT (mô-đun RaBiFPN); **MAE-Base-CP-RaBiFPN** là sự kết hợp giữa bộ giải mã MAE được tiếp tục tiền huấn luyện từ mô hình cơ sở MAE trên ảnh nội soi y tế và bộ giải mã RaBiT (mô-đun RaBiFPN); **MAE-Reg-Base-CP-RaBiFPN** là sự kết hợp giữa bộ giải mã MAE được thêm các **token bổ sung**, được tiếp tục tiền huấn luyện từ mô hình cơ sở trên ảnh nội soi y tế và bộ giải mã RaBiT (mô-đun RaBiFPN).

**Bảng 4.10:** Kết quả thử nghiệm của các mô hình khi huấn luyện trên từng tập dữ liệu với 5% dữ liệu và kiểm thử trên tập kiểm thử tương ứng

Mô hình/Tên bộ dữ liệu	Ung thư dạ dày 5%		Ung thư thực quản 5%		Viêm dạ dày 5%		Viêm thực quản 5%		Viêm loét hành tá tràng 5%		BKAI-Polyp-IGH 5%	
	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU
UNet	82.78	73.1	70.3	58.19	31.84	22.4	42	28.6	45.9	33.22	83.01	75.74
RaBiT	71.33	59.62	53.68	39.22	27.7	18.52	34.22	22.08	3.55	0.8	73.35	64.22
MAE-Scratch-RaBiFPN	74.35	62.92	66.66	53.58	22.95	15.49	35.49	23.23	46.2	33.28	68.92	59.68
MAE-Base-RaBiFPN	78.03	66.89	70.8	59.14	25.2	17.1	39.04	26.18	47.64	34.33	78.07	69.71
MAE-Base-CP-RaBiFPN	82.58	72.76	73.32	61.43	28.66	19.83	40	27.05	48.67	35.1	88.26	82.08
<b>MAE-Reg-Base-CP-RaBiFPN</b>	<b>83.82</b>	<b>74.58</b>	<b>73.9</b>	<b>62.27</b>	<b>34.75</b>	<b>24.58</b>	<b>42.92</b>	<b>29.4</b>	<b>49.02</b>	<b>35.33</b>	<b>89.96</b>	<b>84.37</b>

**Bảng 4.11:** Kết quả thử nghiệm của các mô hình khi huấn luyện trên từng tập dữ liệu với 10% dữ liệu và kiểm thử trên tập kiểm thử tương ứng.

Mô hình/Tên bộ dữ liệu	Ung thư dạ dày 10%		Ung thư thực quản 10%		Viêm dạ dày 10%		Viêm thực quản 10%		Viêm loét hành tá tràng 10%		BKAI-Polyp-IGH 10%	
	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU
UNet	83.86	74.64	72.79	60.94	35.31	25.1	45.1	31.2	46.27	33.68	86.77	80.52
RaBiT	78.46	67.44	64.49	50.96	28.14	17.88	64.49	27.68	29.95	19.55	80.54	72.55
MAE-Scratch-RaBiFPN	78.7	67.71	70.34	57.86	27.54	18.56	40.52	27.39	36.47	25.36	79.33	71.13
MAE-Base-RaBiFPN	81.68	71.75	73.71	62.2	29.28	20.06	43.49	30.12	45.3	32.44	89.27	83.58
MAE-Base-CP-RaBiFPN	84.77	76.01	75.03	63.93	29.58	20.36	45.59	32	51.06	37.33	89.95	84.23
<b>MAE-Reg-Base-CP-RaBiFPN</b>	<b>85.22</b>	<b>76.58</b>	<b>75.41</b>	<b>64.14</b>	<b>35.31</b>	<b>25.18</b>	<b>46.7</b>	<b>32.8</b>	<b>52.49</b>	<b>38.61</b>	<b>90.04</b>	<b>84.51</b>

**Bảng 4.12:** Kết quả thử nghiệm của các mô hình khi huấn luyện trên từng tập dữ liệu và kiểm thử trên tập kiểm thử tương ứng.

Mô hình/Tên bộ dữ liệu	Ung thư dạ dày		Ung thư thực quản		Viêm dạ dày		Viêm thực quản		Viêm loét hành tá tràng		BKAI-Polyp-IGH	
	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU
UNet	86.48	78.58	80.06	70.25	41.43	30.44	56.73	43.22	59.17	45.83	92.37	87.89
RaBiT	83.4	71.21	74.47	64.37	43.68	33.57	53.8	39.88	55.15	40.97	92.44	88.7
MAE-Scratch-RaBiFPN	84.48	75.45	77.92	67.14	36.53	26.48	51.57	37.96	50.48	36.97	91.35	85.98
MAE-Base-RaBiFPN	87.05	79.19	77.66	67.04	46.28	36.17	56.41	42.45	59.38	45.62	93.53	89.18
MAE-Base-CP-RaBiFPN	87.3	79.5	79.14	69.22	48.93	36.87	57.21	43.27	60.5	46.89	93.48	89.15
<b>MAE-Reg-Base-CP-RaBiFPN</b>	<b>88.28</b>	<b>80.83</b>	<b>80.92</b>	<b>71.06</b>	<b>49.12</b>	<b>37.06</b>	<b>59.52</b>	<b>45.49</b>	<b>62.72</b>	<b>48.78</b>	<b>93.7</b>	<b>89.29</b>

Qua các kết quả thử nghiệm ở Bảng 4.10, Bảng 4.11, Bảng 4.12 và Hình 4.4, chúng tôi thấy trong các giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu, mô hình được **tiếp tục** tiền huấn luyện với dữ liệu hình ảnh nội soi y tế cho hiệu suất cao hơn so với mô hình chỉ được huấn luyện với ImageNet-1k. Mô hình được tiền huấn luyện **từ đầu** với dữ liệu hình ảnh nội soi y tế cho hiệu suất kém nhất. Điều này xảy

ra có thể do mô hình được tiền huấn luyện **từ đầu** với hình ảnh nội soi y tế đã chưa thể học, chưa khái quát hóa tốt được những đặc trưng quan trọng của dữ liệu so với việc **tiếp tục** tiền huấn luyện mô hình từ mô hình đã được tiền huấn luyện trên ImageNet-1k khi mà mô hình đã được tiền huấn luyện đó đã học cách khái quát hóa và các đặc trưng của hình ảnh. Đặc biệt, mô hình thêm token bổ sung (register token) cho hiệu suất cao nhất trên toàn bộ tập dữ liệu. Điều đó cho thấy mô hình đã tập trung hơn vào các đối tượng cụ thể, tránh tập trung vào các mảng nền của hình ảnh và học được khả năng tổng quát hóa.

Từ việc thu được các kết quả như trên, chúng tôi có thể kết luận rằng: mô hình đã được tiền huấn luyện trên các bộ dữ liệu lớn như ImageNet-1k nếu được **tiếp tục** huấn luyện sẽ giúp tiết kiệm chi phí, công sức và cũng mang lại hiệu suất cao trong giai đoạn tinh chỉnh cho các nhiệm vụ mục tiêu khi chỉ cần **tiếp tục** tiền huấn luyện trên miền dữ liệu của nhiệm vụ mục tiêu như miền dữ liệu hình ảnh nội soi y tế với ít kỉ nguyên huấn luyện.

Đồng thời, chúng tôi cũng thấy được rằng tuy chỉ sử dụng một lượng ảnh y tế không đủ lớn (khoảng 66000 ảnh) như ImageNet-1k (một triệu ảnh) cho quá trình **tiếp tục** huấn luyện thì mô hình **tiếp tục** huấn luyện đã đạt hiệu suất ấn tượng trong giai đoạn tinh chỉnh. Điều này nói lên rằng trong các miền dữ liệu riêng tư như miền dữ liệu hình ảnh nội soi y tế thì số lượng hình ảnh nội soi y tế được chia sẻ không nhiều, việc tận dụng và **tiếp tục** tiền huấn luyện từ mô hình cơ sở là một giải pháp hiệu quả để tăng hiệu suất cho các nhiệm vụ mục tiêu trong miền dữ liệu riêng tư đó.

#### 4.5.3 Đánh giá trên các điểm chuẩn Polyp

**Bảng 4.13:** Kết quả thử nghiệm của các mô hình trên các điểm chuẩn polyp.

Mô hình/Tên bộ dữ liệu	CVC-300		Kvasir		CVC-ClinicDB		CVC-ColonDB		ETIS-LaribPolypDB	
	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU	Macro Dice	Macro IoU
UNet	74.48	61.37	81.6	72.45	76.02	65.11	61.59	50.3	56.88	46.97
RaBiT	85.27	77.54	88.83	82.5	86.58	80.18	74.53	66.1	64.71	56.13
ColonFormer-S	89.25	82.4	91.44	86.67	93.45	88.44	81.11	73.23	76.86	68.82
ColonFormer-L	90.48	83.98	92.55	87.58	93.7	88.72	80.66	72.83	78.41	70.16
MAE-Base-CP-RaBiFPN	88.4	81.09	89.92	84.84	91.93	87.46	79.44	71.52	77.07	68.85
MAE-Reg-Base-CP-RaBiFPN	88.93	81.54	91.11	85.64	93.9	89.11	80.92	73.06	77.07	69.4

Từ kết quả thử nghiệm ở Bảng 4.13, mô hình phân đoạn của chúng tôi có hiệu suất cao hơn so với RaBiT [20] và có hiệu suất gần ngang bằng với mô hình ColonFormer-S và ColonFormer-L [30]. Các mô hình ColonFormer, RaBiT là những mô hình hiện đạt nhất trên các điểm chuẩn công khai về ảnh nội soi y tế. Việc mô hình có thể gần ngang bằng với các mô hình này chứng tỏ mô hình của chúng tôi có tiềm năng phát triển mạnh mẽ trong tương lai. Đồng thời, mô

hình của chúng tôi cũng có thể được cải thiện bằng cách tinh chỉnh siêu tham số, thêm các phép tăng cường ảnh phù hợp hoặc tiếp tục tiền huấn luyện với lượng lớn ảnh y tế với điều kiện khác nhau như các kích thước ảnh khác nhau, độ sáng khác nhau hoặc thay thế Vision Transformers bằng Mix Transformer (xương sống của ColonFormer) trong MAE rồi tiền huấn luyện tự giám sát với ảnh nội soi y tế.

#### 4.5.4 Mô hình đa nhiệm

Trong phần này, các mô hình của chúng tôi được so sánh với mô hình đa nhiệm EndoUNet [19] trên các nhiệm vụ như phân loại HP, phân loại vị trí giải phẫu, phân loại loại bệnh và phân vùng các tổn thương. Các mô hình đều được huấn luyện trên cùng một tập dữ liệu huấn luyện và được đánh giá trên cùng một tập kiểm tra.

**Bảng 4.14:** Kết quả thử nghiệm phân loại trên các mô hình đa nhiệm.

Mô hình đa nhiệm	Loại bệnh	HP		Vị trí giải phẫu
		Macro Acc	Macro Acc	
EndoUnet (ResNet50)	98.45	89.4	97.57	
MAE-Base-CP-Multitask	97.89	90.45	97.93	
MAE-Reg-Base-CP-Multitask	<b>98.73</b>	<b>92.55</b>	<b>98.47</b>	

**Bảng 4.15:** Kết quả thử nghiệm của các mô hình đa nhiệm trên nhiệm vụ phân vùng

	Ung thư dạ dày		Ung thư thực quản		Viêm dạ dày		Viêm thực quản		Viêm loét hành tá tràng		Polyp	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
EndoUnet (ResNet50)	76.18	84.94	66.74	77.44	29.94	42.17	42.17	56.47	38.86	52.82	86.08	91.2
MAE-Base-CP-Multitask	80.12	87.76	71.68	81.32	37.93	50.4	45.08	59	46.57	60.33	87.24	92.29
MAE-Reg-Base-CP-Multitask	<b>80.68</b>	<b>88.06</b>	<b>72.2</b>	<b>81.69</b>	<b>38.47</b>	<b>50.65</b>	<b>45.51</b>	<b>59.4</b>	<b>46.68</b>	<b>60.41</b>	<b>87.67</b>	<b>92.53</b>

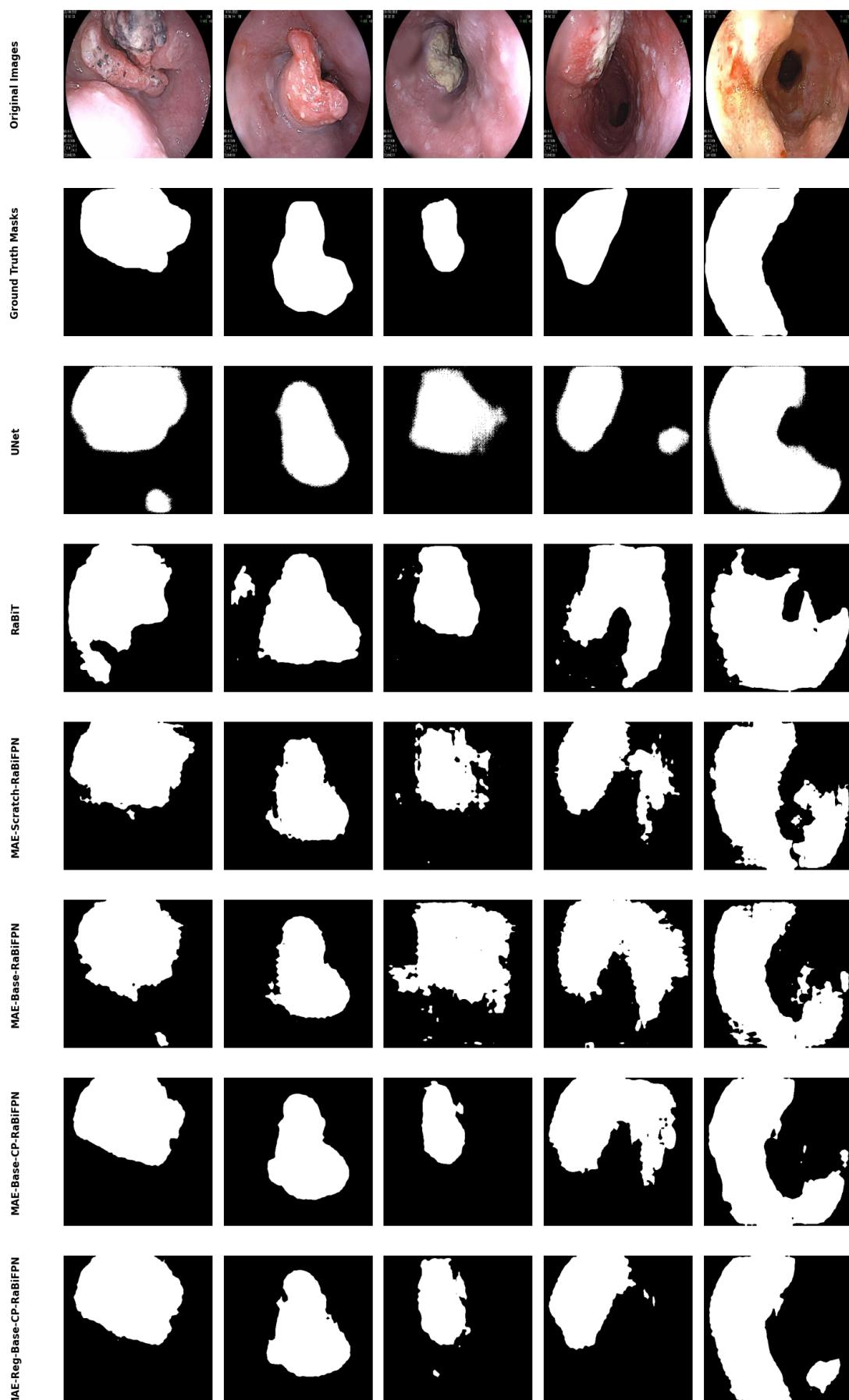
Qua các kết quả thử nghiệm ở Bảng 4.14 và Bảng 4.15, các mô hình đa nhiệm của chúng tôi đều đạt kết quả tốt hơn so với mô hình EndoUNet. Điều này chứng tỏ rằng việc kết hợp nhiều tác vụ giúp mô hình học được nhiều kiến thức hơn, từ đó giúp cải thiện hiệu suất của mô hình. Đồng thời, việc kết hợp nhiều tác vụ cũng giúp giảm thiểu việc overfitting, từ đó giúp mô hình có khả năng tổng quát hóa tốt hơn. Tuy nhiên, việc kết hợp nhiều tác vụ cũng đồng nghĩa với việc tăng thêm chi phí tính toán và tăng thêm thời gian huấn luyện.

Trong tương lai, chúng tôi sẽ tiếp tục thử nghiệm với các tác vụ khác nhau như phát hiện đối tượng, và cố gắng tìm ra cách kết hợp các tác vụ sao cho hiệu quả nhất. Đồng thời, chúng tôi cũng sẽ thử nghiệm với bộ giải mã khác nhau để tìm ra kiến trúc mạng phù hợp nhất cho các tác vụ cụ thể.

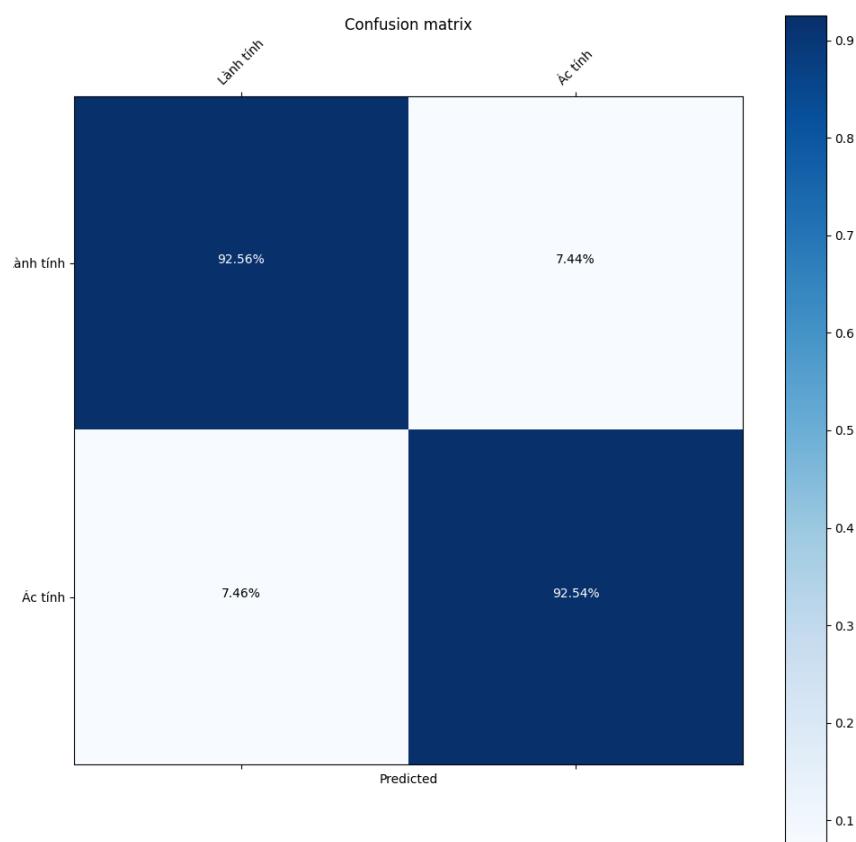
Hình 4.5, Hình 4.6 và Hình 4.7 thể hiện ma trận nhầm lẫn của từng nhiệm vụ

phân loại như phân loại HP, phân loại vị trí giải phẫu và phân loại loại bệnh. Các giá trị trên đường chéo chính thể hiện số lượng các mẫu được phân loại đúng, còn các giá trị còn lại thể hiện số lượng các mẫu bị phân loại nhầm.

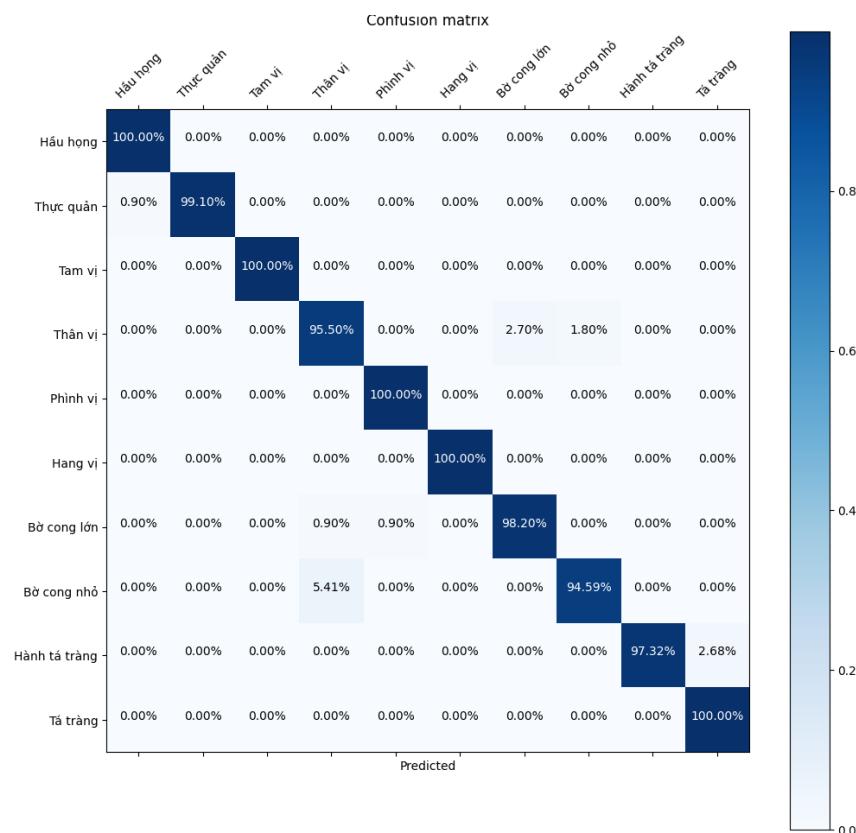
Từ các hình ảnh này, chúng tôi có thể thấy rằng tác vụ phân loại loại bệnh có hiệu suất tốt nhất trong khi tác vụ phân loại HP có hiệu suất thấp nhất. Điều này có thể giải thích bằng việc dữ liệu cho tác vụ phân loại HP khiến mô hình khó phân biệt hơn khi mô hình dễ nhầm lẫn giữa không bị HP và bị HP do vùng bị viêm khá to. Vì vậy, việc phân loại HP cần nhiều dữ liệu hơn và đa dạng hơn để cải thiện hiệu suất.



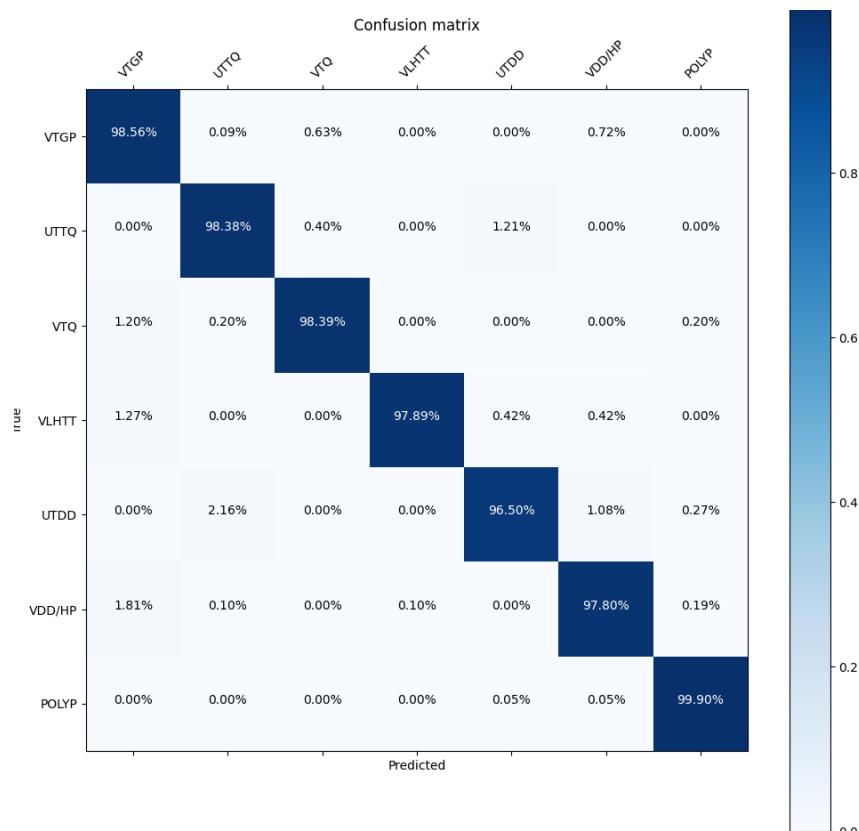
**Hình 4.4:** Một số hình ảnh dự đoán của các mô hình được huấn luyện chỉ với 5% dữ liệu và suy luận trên tập thử nghiệm



**Hình 4.5:** Ma trận nhầm lẫn của phân loại HP sử dụng xương sống của MAE



**Hình 4.6:** Ma trận nhầm lẫn của phân loại vị trí giải phẫu trên mô hình đa nhiệm sử dụng xương sống của MAE



**Hình 4.7:** Ma trận nhầm lẫn của phân loại bệnh trên mô hình đa nhiệm sử dụng xương sống của MAE

## CHƯƠNG 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

### 5.1 Kết luận

Trong đồ án này, chúng tôi đã thử nghiệm phương pháp tự giám sát nhằm tự đào tạo mô hình trên dữ liệu ảnh nội soi tiêu hóa không nhãn với mô hình Masked Autoencoders [4] để tạo ra bộ mã hóa mạnh mẽ làm nền tảng cho việc thực hiện tinh chỉnh trên các nhiệm vụ mục tiêu. Đồng thời, đề xuất một mô hình sử dụng bộ mã hóa của MAE cho nhiệm vụ mục tiêu như phân vùng hình ảnh nội soi y tế và một mô hình đa nhiệm cho hai nhiệm vụ phân vùng và phân loại tổn thương. Qua các kết quả thử nghiệm cho thấy, mô hình của chúng tôi đạt được hiệu suất tương đương và cao hơn các mô hình cơ sở trong các trường hợp. Do đó, chúng tôi tin rằng mô hình của chúng tôi có thể được ứng dụng trong thực tế để hỗ trợ bác sĩ trong việc phân tích hình ảnh nội soi y tế.

### 5.2 Hướng phát triển trong tương lai

Trong tương lai, đồ án có thể cải tiến và phát triển thêm nhằm nâng cao chất lượng và hiệu suất hơn nữa. Thứ nhất, cần bổ sung thêm lượng lớn dữ liệu hình ảnh nội soi y tế không gán nhãn cho quá trình tiền huấn luyện. Những dữ liệu này phải thật phong phú về vị trí trong các cơ quan tiêu hóa cũng như các loại bệnh khác nhau. Thứ hai, cần phải tăng cường thêm các kỹ thuật tiền xử lý dữ liệu như chuẩn hóa, cân bằng dữ liệu, tăng cường dữ liệu, loại bỏ nhiễu, ... để cải thiện chất lượng dữ liệu đầu vào cho quá trình tiền huấn luyện và quá trình tinh chỉnh cho các nhiệm vụ mục tiêu. Thứ ba, có thể cải thiện mạng xương sống bằng cách sử dụng mô hình mới ra mắt gần đây là Hiera [31]. Đây cũng là một mô hình che giấu hình ảnh sử dụng phương pháp giống như Masked Autoencoders [4]. Mô hình này có cải tiến kiến trúc của Vision Transformer [9] thành kiến trúc có hình dạng giống như ResNet [13]. Điều này làm cho mô hình đơn giản hơn, học được các biểu diễn hiệu quả hơn so với Vision Transformer, thậm chí mô hình còn suy luận nhanh hơn và đạt được hiệu suất cao.

## TÀI LIỆU THAM KHẢO

- [1] H. Bao, L. Dong, S. Piao, and F. Wei, “Beit: BERT pre-training of image transformers,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.
- [2] J. Zhou, C. Wei, H. Wang, *et al.*, “Image BERT pre-training with online tokenizer,” in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, OpenReview.net, 2022.
- [3] Z. Xie, Z. Zhang, Y. Cao, *et al.*, “Simmim: A simple framework for masked image modeling,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, IEEE, 2022, pp. 9643–9653.
- [4] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick, “Masked autoencoders are scalable vision learners,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, IEEE, 2022, pp. 15 979–15 988.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1106–1114.
- [6] W. Januszewicz, K. Witczak, P. Wieszczy, *et al.*, “Prevalence and risk factors of upper gastrointestinal cancers missed during endoscopy: A nationwide registry-based study,” *Endoscopy*, vol. 54, no. 07, pp. 653–660, 2022.
- [7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, IEEE Computer Society, 2009, pp. 248–255.
- [8] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, *et al.*, Eds., 2017, pp. 5998–6008.

- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.
- [10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [11] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 1597–1607.
- [12] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. Cambridge University Press, 2023, <https://D2L.ai>.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, IEEE Computer Society, 2016, pp. 770–778.
- [14] L. J. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *CoRR*, vol. abs/1607.06450, 2016.
- [15] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, ser. JMLR Workshop and Conference Proceedings, vol. 28, JMLR.org, 2013, pp. 1310–1318.
- [16] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, N. Navab, J. Hornegger, W. M. W. III, and A. F. Frangi, Eds., ser. Lecture Notes in Computer Science, vol. 9351, Springer, 2015, pp. 234–241.
- [17] D. Fan, G. Ji, T. Zhou, *et al.*, “Pranet: Parallel reverse attention network for polyp segmentation,” in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VI*, A. L. Martel, P. Abolmaesumi, D. Stoyanov, *et al.*, Eds., ser. Lecture Notes in Computer Science, vol. 12266, Springer, 2020, pp. 263–273.

- [18] S. Chen, X. Tan, B. Wang, and X. Hu, “Reverse attention for salient object detection,” in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., ser. Lecture Notes in Computer Science, vol. 11213, Springer, 2018, pp. 236–252.
- [19] N. D. Manh, D. V. Hang, D. V. Long, *et al.*, “Endounet: A unified model for anatomical site classification, lesion categorization and segmentation for upper gastrointestinal endoscopy,” in *14th International Conference on Knowledge and Systems Engineering, KSE 2022, Nha Trang, Vietnam, October 19-21, 2022*, N. B. Hung, L. S. Vinh, N. L. Minh, *et al.*, IEEE, 2022, pp. 1–6.
- [20] N. H. Thuan, N. T. Oanh, N. T. Thuy, S. W. Perry, and D. V. Sang, “Rabit: An efficient transformer using bidirectional feature pyramid network with reverse attention for colon polyp segmentation,” *CoRR*, vol. abs/2307.06420, 2023.
- [21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 10347–10357.
- [22] A. Radford, J. W. Kim, C. Hallacy, *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8748–8763.
- [23] M. Oquab, T. Darcet, T. Moutakanni, *et al.*, “Dinov2: Learning robust visual features without supervision,” *CoRR*, vol. abs/2304.07193, 2023.
- [24] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, “Vision transformers need registers,” in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=2dnO3LLiJ1>.
- [25] M. Tan, R. Pang, and Q. V. Le, “Efficientdet: Scalable and efficient object detection,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, Computer Vision Foundation / IEEE, 2020, pp. 10778–10787.

- [26] C. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, “Deeply-supervised nets,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, G. Lebanon and S. V. N. Vishwanathan, Eds., ser. JMLR Workshop and Conference Proceedings, vol. 38, JMLR.org, 2015.
- [27] G. Jocher, A. Chaurasia, and J. Qiu, *Ultralytics YOLO*, version 8.0.0, Jan. 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>.
- [28] A. Hatamizadeh, Y. Tang, V. Nath, *et al.*, “UNETR: transformers for 3d medical image segmentation,” in *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, IEEE, 2022, pp. 1748–1758.
- [29] Z. Dou, Y. Xu, Z. Gan, *et al.*, “An empirical study of training end-to-end vision-and-language transformers,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, IEEE, 2022, pp. 18 145–18 155.
- [30] T. D. Nguyen, T. Nguyen, N. T. Thuy, M. Tran, and D. V. Sang, “Colonformer: An efficient transformer based method for colon polyp segmentation,” *IEEE Access*, vol. 10, pp. 80 575–80 586, 2022.
- [31] C. Ryali, Y. Hu, D. Bolya, *et al.*, “Hiera: A hierarchical vision transformer without the bells-and-whistles,” in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 29 441–29 454.