TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Thái Bảo - Trần Thị Thanh Phú

XÁC ĐỊNH TẾ BÀO UNG THƯ TUYẾN TIỀN LIỆT TỪ HÌNH ẢNH BỆNH LÝ SỬ DỤNG MẠNG PHÂN TÍCH KIM TỰ THÁP

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

Tp. Hồ Chí Minh, tháng 03/2024

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

Nguyễn Thái Bảo - 20127448
Trần Thị Thanh Phú - 20127279

XÁC ĐỊNH TẾ BÀO UNG THƯ TUYẾN TIỀN LIỆT TỪ HÌNH ẢNH BỆNH LÝ SỬ DỤNG MẠNG PHÂN TÍCH KIM TỰ THÁP

KHÓA LUẬN TỐT NGHIỆP CỬ NHÂN
CHƯƠNG TRÌNH CHẤT LƯỢNG CAO

GIÁO VIÊN HƯỚNG DẪN
PGS.TS. Lê Hoàng Thái

Tp. Hồ Chí Minh, tháng 03/2024

Lời cảm ơn

Lời đầu tiên, nhóm chúng em muốn bày tỏ lòng biết ơn sâu sắc đến PGS.TS Lê Hoàng Thái - người thầy đã dành thời gian và tâm huyết để hướng dẫn, chia sẻ kiến thức và kinh nghiệm quý báu cho chúng em trong suốt hành trình thực hiện khoá luận này. Nhờ có sự hướng dẫn tận tình của thầy, chúng em đã có thể vượt qua được những khó khăn, giải quyết các vấn đề một cách khoa học và hoàn thiện nhất.

Nhóm chúng em cũng muốn bày tỏ lòng biết ơn đến Khoa Công nghệ thông tin, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM đã tạo điều kiện và môi trường học tập và nghiên cứu rất tốt cho sinh viên. Nhờ vào sự đào tạo chuyên sâu và những kiến thức hữu ích mà chúng em có thể áp dụng linh hoạt và thành công trong quá trình hoàn thiện khoá luận này.

Nhóm chúng em thừa nhận rằng, do thiếu sót về kinh nghiệm cũng như kiến thức, khoá luận của chúng em không tránh khỏi những hạn chế. Chính vì vậy, chúng em rất mong nhận được sự đánh giá, ý kiến đóng góp và phê bình từ phía quý thầy cô để chúng em có thể hoàn thiện và nâng cao chất lượng khoá luận một cách tốt nhất.

Cuối cùng, nhóm chúng em xin gửi lời chúc tới quý thầy cô, chúc quý thầy cô luôn dồi dào sức khỏe, thành công rực rỡ trong sự nghiệp giảng dạy và cuộc sống. Chân thành cảm ơn quý thầy cô đã luôn bên cạnh, động viên và hỗ trợ chúng em trong suốt thời gian qua.

TP. Hồ Chí Minh, ngày 12 tháng 02 năm 2024

Nhóm sinh viên thực hiện

Nguyễn Thái Bảo - Trần Thị Thanh Phú

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ CƯƠNG KHOÁ LUẬN TỐT NGHIỆP

XÁC ĐỊNH TẾ BÀO UNG THƯ TUYẾN TIỀN LIỆT TỪ HÌNH ẢNH BỆNH LÝ SỬ DỤNG MẠNG PHÂN TÍCH KIM TỰ THÁP

(Identyfy Prostate Cancer from Pathology Images using Pyramid Scene Parsing Network)

1. THÔNG TIN CHUNG

Người hướng dẫn:

- PGS.TS. Lê Hoàng Thái.

[Nhóm]Sinh viên thực hiện:

1. Nguyễn Thái Bảo (MSSV : 20127448).
2. Trần Thị Thanh Phú (MSSV: 20127279)

Loại đề tài: Nghiên cứu.

Thời gian thực hiện: Từ *tháng 09/2023* đến *tháng 03/2024*.

2. NỘI DUNG CHI TIẾT

2.1. Giới thiệu đề tài

Ung thư tuyến tiền liệt (PCa) là căn bệnh thứ sáu phổ biến nhất và gây tử vong thứ hai nhiều nhất ở nam giới trên toàn thế giới. Có nhiều kỹ thuật khác nhau để phát hiện và đánh giá PCa. Tuy nhiên, việc kiểm tra siêu vi kính của mẫu mô sinh thiết đã được nhuộm màu bởi các bác sĩ bệnh lý là phương pháp chính xác nhất. Dựa trên các mẫu histological có thể quan sát được, từng vùng của mô được gán một Gleason grade từ 1 đến 5. Điểm Gleason cuối cùng được báo cáo là tổng của hai mẫu rõ nét nhất và mẫu thứ hai rõ nét nhất; ví dụ, một mẫu mô có mẫu rõ nét nhất là Gleason grade 4 và mẫu thứ hai rõ nét nhất là Gleason grade 3 sẽ có điểm Gleason là 4+3.

Việc đánh giá Gleason của ung thư tuyến tiền liệt thường được thực hiện thông qua kiểm tra bằng mắt (bằng kính hiển vi) mẫu mô tuyến tiền liệt bởi các bác sĩ bệnh lý chuyên gia. Tuy nhiên, đây là một nhiệm vụ tốn thời gian và có sự biến động lớn giữa các bác sĩ, vì vậy các phương pháp tự động hỗ trợ máy tính có tiềm năng cải thiện tốc độ, độ chính xác và khả năng tái lập của kết quả.

2.2. Mục tiêu đề tài

Với sự phát triển của lĩnh vực học sâu, rất nhiều mô hình học sâu áp dụng cho bài toán phân đoạn ảnh đã suất hiện. Trong vài năm trở lại đây, các mô hình đã được áp dụng nhiều hơn trong lĩnh vực y học để thực hiện các tác vụ phân đoạn các tế bào bất thường trong hình ảnh bệnh lý. Mặc dù độ chính xác của các mô hình ngày càng được cải thiện nhưng việc đưa vào thực tiễn vẫn còn gặp nhiều thách thức to lớn.

Với đề tài “Xác định tế bào ung thư tuyến tiền liệt từ hình ảnh bệnh lý” đã đặt mục tiêu quan trọng là phát triển và ứng dụng các giải pháp tự động hóa trong việc đưa ra dự đoán về các phân vùng bị ung thư từ hình ảnh bệnh lý. Thách thức phản ảnh sự cần thiết

của việc sử dụng trí tuệ nhân tạo để cải thiện quá trình chuẩn đoán và dự đoán Pca, một căn bệnh quan trọng và phổ biến ảnh hưởng đến nam giới trên toàn thế giới.

Mục tiêu cuối cùng là giúp cải thiện tốc độ, độ chính xác và tính đáng tin cậy của phân vùng ung thư, từ đó tạo điều kiện thuận lợi cho các bác sĩ trong việc đưa ra quyết định điều trị và dự đoán tiến triển của bệnh cho bệnh nhân.

2.3. Phạm vi của đề tài

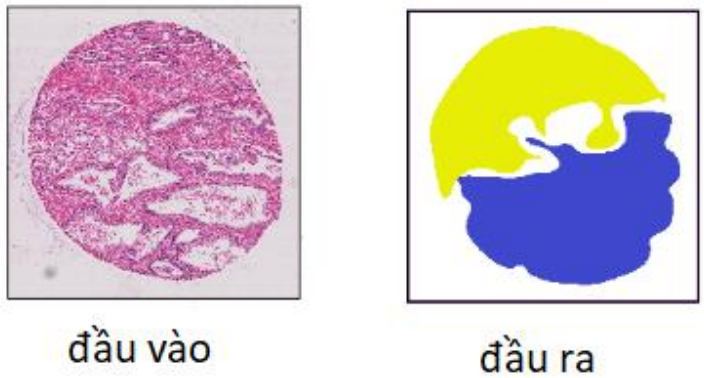
Nội dung nghiên cứu chính của đề tài là tập trung nghiên cứu mô hình học sâu đã đạt kết quả khá tốt gần đây là mô hình PSPNet [1]. Đầu tiên, nhóm sẽ tìm hiểu lý thuyết, thực nghiệm, phân tích các ưu điểm và khuyết điểm của mô hình. Từ đó nhóm sẽ nghiên cứu và đề xuất một số phương pháp để cải tiến mô hình.

Quá trình trên thực hiện dựa trên tập dữ liệu Gleason là: Gleason 2019 Challenge [2].

2.4. Cách tiếp cận dự kiến

2.4.1. Bài toán

Bài toán xác định phân vùng tế bào ung thư tuyến tiền liệt từ hình ảnh bệnh lý có đầu vào là các hình ảnh bệnh lý (Histopathology) về các tế bào và đầu ra của bài toán là phân vùng các điểm ung thư trên tế bào với các giai đoạn. Để thực hiện bài toán này, mô hình học sâu sẽ phải phân loại từng điểm ảnh của hình ảnh đầu vào thuộc vào phân loại các lớp là các giai đoạn ung thư (Gleason grade) gồm các lớp “Begin”, “Gleason grade 3”, “Gleason grade 4”, “Gleason grade 5”. Hình 1 là ví dụ của đầu vào và đầu ra của bài toán.



Hình 1: Hình bên trái là hình ảnh tế bào được đưa vào mô hình để dự đoán. Hình bên phải là đầu ra của mô hình với các phân vùng bị ung thư với các giai đoạn

Các độ đo thường dùng cho bài toán liên quan đến phân đoạn ảnh bao gồm: Accuracy, Precision, Recall, IoU (Intersection over Union), Dice, v.v… . Trong đề tài này, nhóm quyết định sử dụng độ đo mean IoU (trung bình điểm IoU của các ảnh đầu vào) và độ đo Dice (được biết đến là F1 Score) để đánh giá kết quả thực nghiệm của mô hình. Công thức của 2 độ đo được thể hiện ở dưới hình 2.

$$\text{IoU} = \frac{Area\ of\ overlap}{Area\ of\ union} =$$

$$\text{Dice} = \frac{2\ X\ Area\ of\ overlap}{Total\ area} = 2\ x$$

Hình 2: Công thức của độ đo Iou và Dice được sử dụng trong việc đánh kết quả đầu ra của bài toán

2.4.2. Các công trình liên quan

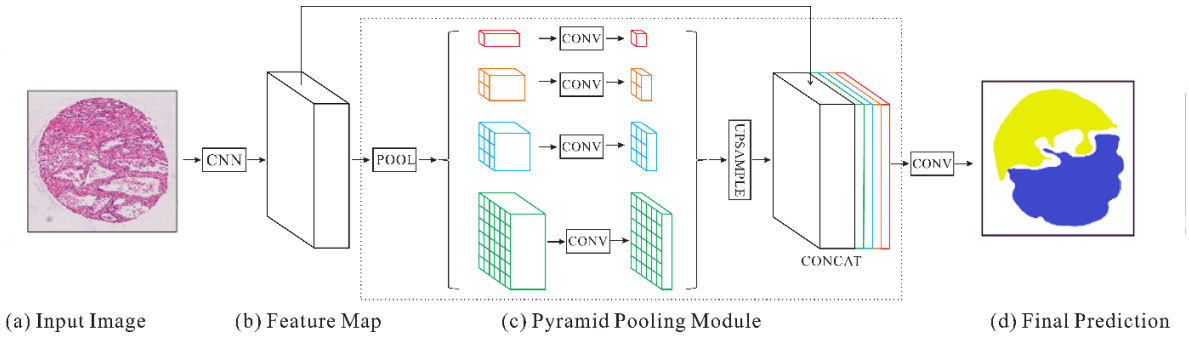
Nhiều phương pháp ứng dụng mạng học sâu đã được nghiên cứu và phát triển để thực hiện bài toán phân đoạn ảnh bệnh lý vào những năm gần đây. Đầu tiên có thể kể đến U-Net [3]. Về cơ bản, U-Net [3] sử dụng kiến trúc mã hóa – giải mã (encoder –

decoder), bao gồm một đường thu hẹp để trích xuất đặc trưng và một đường mở rộng với một lớp phân loại, để thực hiện phân loại nhị phân đối với từng điểm ảnh trong một bản đồ đặc trưng đã được lấy mẫu. Bên cạnh đó có FCN [4], là một kiến trúc được sử dụng chủ yếu để phân đoạn theo ngữ nghĩa. Chúng chỉ sử dụng các lớp được kết nối cục bộ, chẳng hạn như tích chập, gộp và lấy mẫu lại. DeepLab [5], là một kiến trúc áp dụng một cách rất linh hoạt tích chập Atrous thay vì các phương pháp trước đó là áp dụng Transposed Convolution. Bên cạnh đó tác giả cũng áp dụng phương pháp Conditional Random Field để tinh chỉnh kết quả dự báo chuẩn xác hơn.

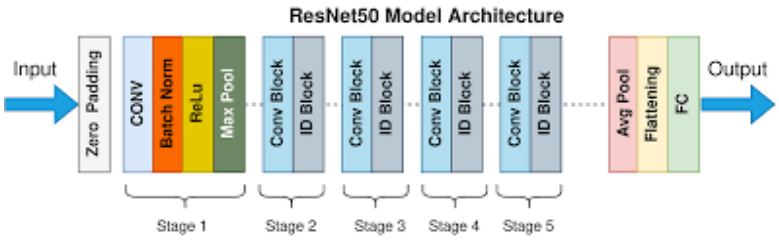
Mặc dù có rất nhiều mô hình đã được đề xuất và đạt kết quả khá tốt cho bài toán phân đoạn hình ảnh bệnh lý nhưng vẫn chưa có nhiều thực nghiệm để kiểm chứng độ hiệu quả của những mô hình đó cho bài toán phân đoạn tế bào ung thư nhóm sẽ thực hiện cải tiến dựa trên mô hình PSPNet [1].

2.4.3. Mô hình

PSPNet [1] là một mô hình phân đoạn ngữ nghĩa sử dụng mô-đun phân tích cú pháp kim tự tháp để khai thác thông tin ngữ cảnh toàn cầu bằng cách tổng hợp ngữ cảnh dựa trên các khu vực khác nhau. Sử dụng mạng CNN đã được huấn luyện từ trước(Resnet [6]) để đưa ra được bản đồ đặc trưng (features map) có kích thước bằng 1/8 so với ảnh gốc. Sau đó đi qua một mô-đun tổng hợp kim tự tháp với 4 cấp độ khác nhau sau đó được hợp nhất lại với nhau và thực hiện việc upsampling để đưa về kích thước bằng với kích thước ảnh gốc.



Hình 3: Mô hình PSPNet [1]



Hình 4: Kiến trúc mạng Resnet [6]

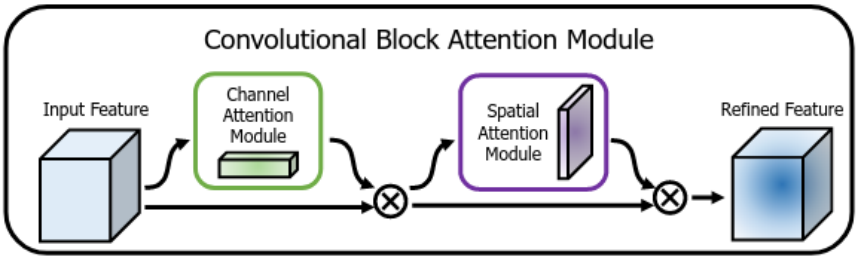
Mục tiêu chính của mô-đun tổng hợp kim tự tháp (pyramid pooling module) là nâng cao khả năng phân đoạn các vùng khác nhau trong hình ảnh với các kích thước vùng quan sát khác nhau. PPM hoạt động bằng cách chia hình ảnh đầu vào thành các phạm vi hoặc cửa sổ con với các kích thước khác nhau, sau đó áp dụng một phép gom nhóm (pooling operation) trong từng cửa sổ con để trích xuất thông tin cụ thể từ mỗi phạm vi. Sau đó, các thông tin từ các cửa sổ con này được tổng hợp lại để tạo ra một biểu đồ của hình ảnh ở nhiều tỷ lệ kích thước khác nhau. Điều này giúp mô hình có khả năng phân đoạn các vật thể ở các kích thước và tỷ lệ khác nhau trong hình ảnh.

2.4.4. Hướng dẫn cải tiến đề xuất

a. Đề xuất một số cơ chế chú ý để cải thiện thông tin tích chập.

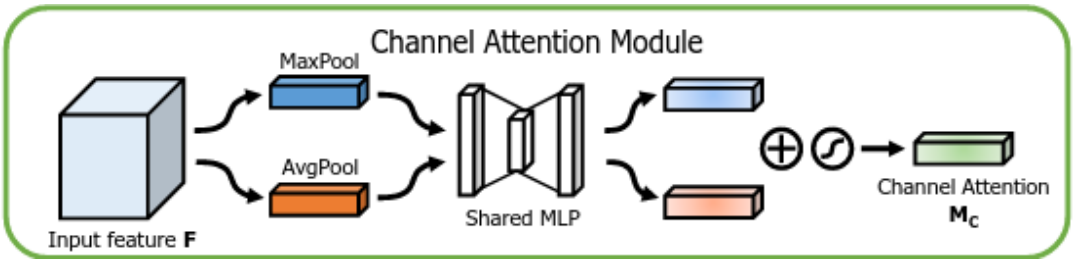
- CBAM: Convolutional Block Attention Module

CBAM [7] bao gồm 2 phần là Channel Attention và Spatial Attention. Input feature maps sẽ được thực hiện Channel Attention trước, sau đó thực hiện tiếp Spatial Attention (Hình 5).



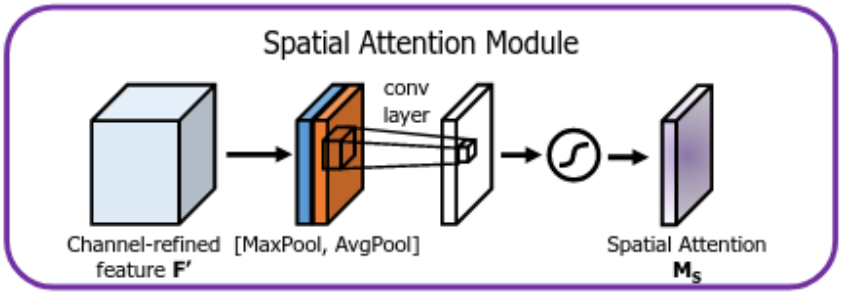
Hình 5: Khối CBAM [7]

Channel Attention (Hình 6). Gồm phần Squeeze (thu) và phần Excitation (phóng). Phần Squeeze có trách nhiệm thu gom thông tin toàn cục (global information) sử dụng Global Average Pooling (GAP) và GMP (Global Max Pooling). Phần Excitation có trách nhiệm tạo attention trên chiều channel sử dụng lớp Fully Connected (FC) với activation function. Tác giả của CBAM có nói rằng GMP cũng thu thập thông tin quan trọng của vật thể nhưng theo một khía cạnh khác. Sau đó, features thu được thì GMP và GAP đều được cho đi qua chung phần Excitation (phóng).



Hình 6: Channel Attention trong CBAM

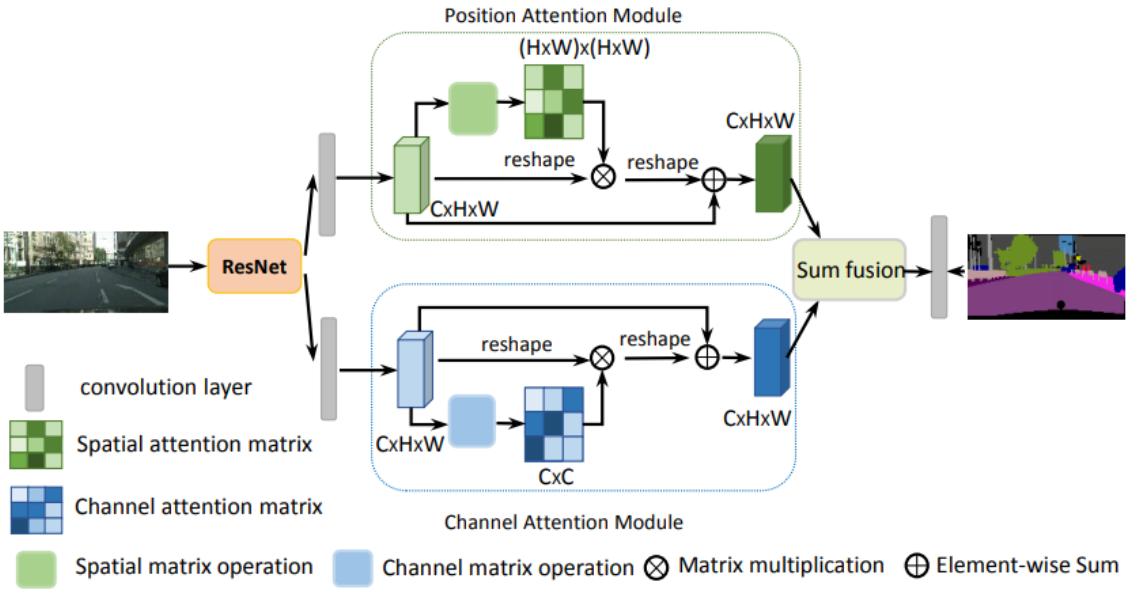
Spatial Attention (Hình 7). Cách thực hiện Spatial Attention của CBAM rất đơn giản, tương tự như Channel Attention. Thay vì thực hiện GAP và GMP theo chiều channel, ta thực hiện GAP và GMP theo chiều spatial của feature maps, tạo ra 2 feature maps tổng hợp với số chiều channel là 1. Hai feature maps này sau đó được nối lại với nhau và đi qua một 7×7 Conv kèm Sigmoid với số filter là 1 để tạo ra attention. Đặc biệt là phải dùng một Conv có kernel size to (chẳng hạn như 7x7) để có thể tính toán được vùng thông tin rộng hơn và tổng quát hơn.



Hình 7: Spatial Attention trong CBAM

– Dual Attention Network

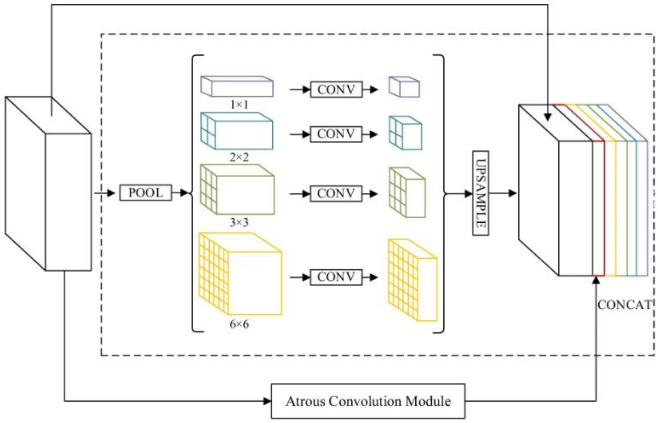
Dual attention netword [8] được đề xuất để tích hợp các đặc trưng cục bộ với sự phụ thuộc toàn cục của chúng. DAN bao gồm 2 thành phần chính là Positional Attention Module (PAM) và Channel Attention Module (CAM). Trong đó, PAM tổng hợp có chọn lọc những đặc trưng tại từng vị trí bằng tổng trọng số của đặc trưng ở tất cả vị trí. Những đặc trưng tương tự sẽ liên quan đến nhau bất kể khoảng cách giữa chúng. Trong khi đó, CAM nhấn mạnh có chọn lọc sự phụ thuộc giữa các kênh bằng cách tích hợp những đặc trưng liên quan giữa chúng. Đầu ra của hai khối này sẽ được đi qua một hàm được gọi là Sum fusion để tổng hợp sự biểu diễn các đặc trưng, góp phần nâng cao hiệu quả phân đoạn. (Hình 8).



Hình 8: Kiến trúc mô hình Dual Attention Network [8]

b. Thêm vào một nhánh mô-đun tích chập astrous (Astrous convolution module) vào mô-đun kim tự tháp (Pyramid pooling module) nhằm cải thiện việc thu thập thông tin của mô hình.

Việc thêm một nhánh astrous convolution [9] vào mô-đun kim tự tháp nhằm trích xuất các đặc điểm chung từ các đặc điểm nông trong quá trình lan truyền qua mạng cnn, đồng thời mở rộng cùng receptive fields. Và đầu ra của nhánh astrous convolution là một vùng thông tin nông thu thập từ ảnh đầu vào. Đầu ra của mô-đum kim tự tháp được kết hợp với đầu ra của nhánh astrous convolution để mạng có thể giữ được thông tin toàn cục vào cục bộ.



Hình 9: Mô-đun kim tự tháp (Pyramid pooling module) kết hợp với astrous convolution

c. *Sử dụng mô hình đã được huấn luyện từ trước như Xception , … để nâng cao hiệu quả huấn luyện.*

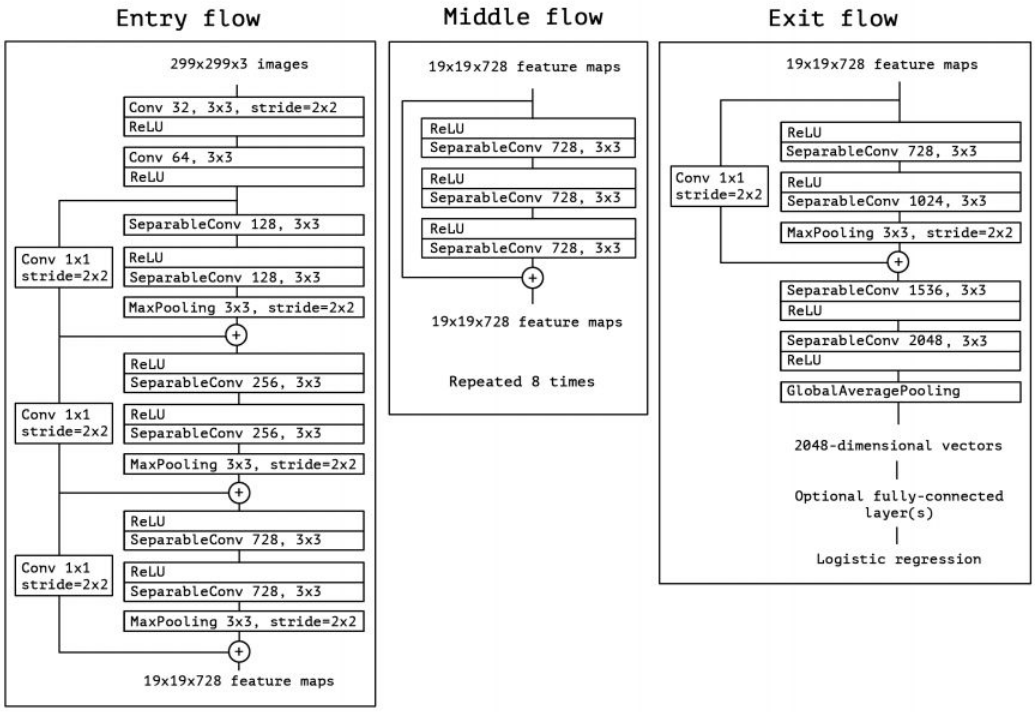
Xception [10] là một kiến trúc mạng tích chập sâu bao gồm các kết cấu có thể phân tách theo chiều sâu. Nó được phát triển bởi các nhà nghiên cứu của Google. Google đã trình bày cách diễn giải các mô-đun Inception trong mạng nơ ron tích chập như một bước trung gian giữa phép tích chập thông thường và phép toán tích chập có thể phân tách theo chiều sâu (một phép tích chập theo chiều sâu, sau đó là một phép tích chập theo điểm).

Xception có cấu trúc dựa trên hai điểm chính:

- Tích chập có thể phân tách theo chiều sâu (Depthwise Separable Convolution).

- Các kết nối tương tự như resitdual của resnet được gọi là phím tắt giữa các khối (Shortcuts between Convolution blocks).

Xception gồm 3 khối lớn chính là Entry flow, Middle flow, Exit flow.



Hình 10: kiến trúc Xception [10]

Việc nhóm sự dụng kiến trúc đã huấn luyện xception vì theo kết quả đánh giá của tác giả cho thấy Xception [10] sử dụng kiến trúc tích chập phân tách theo chiều sâu (Convolutional depthwise separable) thay vì convolution thông thường như Resnet [6]. Kiến trúc này giúp mô hình có thể học được các đặc trưng ở mức độ chi tiết cao hơn và giảm lượng tham số cần thiết.

d. Kĩ thuật test time augmentation (TTA).

Kĩ thuật Test time agumentation [11] là một kĩ thuật được dùng trong việc đánh giá mô hình trong quá trình huấn luyện, việc thực hiện các biến đổi phù hợp đối với tập dữ liệu

kiểm tra nhằm cải thiện hiệu suất dự đoán tổng thể. Cụ thể trong TTA [11], kỹ thuật tăng cường dữ liệu sẽ được áp dụng trên từng ảnh của tập kiểm tra, nhiều ảnh tăng cường mới sẽ được tạo ra. Mô hình sẽ thực hiện dự đoán riêng trên các ảnh tăng cường này và trả về kết quả dự đoán trung bình trên các cách biến đổi ảnh.

Nhóm sẽ áp dụng kĩ thuật này theo hai hướng là lật dọc (vertical flip) và lật ngang (horizontal flip) cho quá trình đánh giá mô hình.

2.5. Kết quả dự kiến của đề tài

Kết quả dự kiến của đề tài bao gôm:

- Hiểu được tổng quan bài toán phân đoạn từ hình ảnh bệnh lý và một số mô hình tiêu biểu trong lĩnh vực này.
- Nắm được cơ chế hoạt động, điểm mạnh và yếu của mô hình PSPNet. Nhóm sẽ tiến hành huấn luyện mô hình lại từ đầu trên tập dữ liệu đã nêu trên
- Kết hợp các phương pháp, kĩ thuật cải tiến đã nêu trên và tiến hành đánh giá tường hình thước cải tiến. Sau đó lập bảng đánh giá so sánh kết quả đạt được để đưa ra được phương pháp cải tiến tối ưu nhất.
- Cuối cùng là so sánh kết quả của nhóm với các mô hình khác.

2.6. Kế hoạch thực hiện

THỜI GIAN	CÔNG VIỆC	PHÂN CÔNG
Tháng 8/2023	1. Nghiên cứu các khái niệm và lý thuyết cho bài toán phân đoạn ảnh bệnh lý nói chung và bài toán phân đoạn tế bào ung thư nói riêng. 2. Tìm hiểu về các công trình nghiên cứu đã được đề xuất để giải quyết bài toán.	Tất cả thành viên
Tháng 9/2023	1. Thảo luận và chọn ra một mô hình tiêu biểu để nghiên cứu sâu.	Tất cả thành viên

	2. Nghiên cứu chi tiết về bài báo của mô hình đã chọn. 3. Xử lý các tập dữ liệu phù hợp cho việc thực nghiệm của mô hình.	
Tháng 10/2023	1. Đọc hiểu mã nguồn PSPNet của bài báo gốc 2. Tiến hành cài đặt trên máy cá nhân để thực nghiệm	Thái Bảo
Tháng 11/2023	1. Từ kết quả thực nghiệm, phân tích cơ chế hoạt động, các tính chất, điểm mạnh và điểm yếu của mô hình. 2. Đề xuất hướng cải tiến cho mô hình	Tất cả thành viên
Tháng 12/2023	1. Tiến hành cài đặt các hướng cải tiến dựa trên mô hình gốc. 2. Huấn luyện mô hình cải tiến dựa trên tập dữ liệu đã xử lý . 3. Lập bảng so sánh đánh giá độ hiệu quả, thời gian huấn luyện. 4. Chọn ra mô hình cải tiến tối ưu nhất và hiệu quả nhất để thực hiện huấn luyện và đánh giá chi tiết nhất.	Tất cả thành viên
Tháng 2-3/2024	1. Viết báo cáo khóa luận 2. Chuẩn bị powerpoint thuyết trình	Tất cả thành viên

Tài liệu tham khảo

[1] J. S. X. Q. X. W. J. J. Hengshuang Zhao, Pyramid Scene Parsing Network, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[2] Danielle Walker. Miccai automatic prostate gleason grading challenge, 2019.

[3] O. RONNEBERGER, P. FISCHER and T. BROX, U-net: Convolutional networks for biomedical image segmentation, International Conference on

Medical image computing and computer-assisted intervention. Springer, Cham, 2015. p. 234-241..

[4] E. S. a. T. D. J. Long, Fully convolutional networks for semantic segmentation, In CVPR, 2015.

[5] G. P. I. K. K. M. a. A. L. Y. L. Chen, Semantic image segmentation with deep convolutional nets and fully connected, crfs. arXiv:1412.7062, 2014..

[6] X. Z. S. R. a. J. S. K. He, Deep residual learning for image recognition., In CVPR, 2016.

[7] J. P. J.-Y. L. I. S. K. Sanghyun Woo, CBAM: Convolutional Block Attention Module, In CVPR, [v2] 2018.

[8] J. e. a. FU, Dual attention network for scene segmentation., In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 3146-3154..

[9] "A Primer on Atrous(Dilated) and Depth-wise Separable Convolutions," [Online]. Available: https://towardsdatascience.com/a-primer-on-atrous-convolutions-and-depth-wise-separable-convolutions-443b106919f5.

[10] F. Chollet., Xception: Deep Learning with Depthwise Separable Convolutions., In CVPR [v3] 2017..

[11] M. KIMURA, Understanding Test-Time Augmentation., In International Conference on Neural Information Processing. Springer, Cham, 2021. p. 558-569.F.

TP. Hồ Chí Minh, ngày ... tháng ... năm .....

XÁC NHẬN CỦA NGƯỜI HƯỚNG DẪN	**NHÓM SINH VIÊN THỰC HIỆN**
(Ký và ghi rõ Họ tên)	**(Ký và ghi rõ Họ tên)**

Mục lục

Lời cảm ơn **i**

Đề cương chi tiết

Mục lục **ii**

Tóm tắt **vi**

1 GIỚI THIỆU ĐỀ TÀI **1**

1.1 Đặt vấn đề 1

1.2 Mục tiêu nghiên cứu 1

1.3 Phạm vi đề tài 2

1.4 Đóng góp của đề tài 2

1.5 Cấu trúc khóa luận 3

2 CÁC CÔNG TRÌNH LIÊN QUAN **4**

2.1 U-Net [31] 4

2.1.1 Kiến trúc mô hình 4

2.1.2 Kết quả thực nghiệm 6

2.2 FCN - Fully Convolutional Networks [26] 7

2.2.1 Kiến trúc mô hình 7

2.2.2 Kết quả thực nghiệm 8

2.3 Các thuật toán liên quan 10

2.3.1 Majority Voting Algorithm [34] 11

2.3.2 STAPLE Algorithm [34] 12

2.4 PSPNet - Pyramid Scene Parsing Network [41] 15

2.4.1 Kiến trúc mô hình 15

2.4.2 Kết quả thực nghiệm 16

2.5 Tổng kết 18

3 PHƯƠNG PHÁP ĐỀ XUẤT **20**
3.1 Đề xuất một số kiến trúc có thể được kết hợp 20
3.1.1 Convolutional Block Attention Module 20
3.1.2 Atrous convolution Spatial Pyramid 23
3.1.3 Global Attention Upsample[22] 28
3.2 Kiến trúc đề xuất 30
3.2.1 Mô-đun đề xuất 30
3.2.2 Kiến trúc đề xuất 32
3.2.3 Mạng lưới chi nhánh phụ trợ(Auxiliary Branch Network)[11] 35
3.2.4 Tổng kết 36

4 CÀI ĐẶT VÀ THỰC NGHIỆM **37**
4.1 Chuẩn bị dữ liệu 37
4.2 Chiến lược đánh giá 39
4.3 Môi trường thực nghiệm 41
4.4 Cài đặt thực nghiệm 41
4.4.1 Ngôn ngữ và thư viện lập trình 41
4.4.2 Tiền xử lý và tăng cường dữ liệu 42
4.4.3 Các tham số và kỹ thuật thực nghiệm 43
4.5 Quy trình thực nghiệm 44
4.6 Kết quả và nhận xét 44
4.6.1 Mô hình PSPNet[41] 44
4.6.2 Mô hình đề xuất 45
4.7 Đánh giá kết quả 47

5 Kết luận và hướng phát triển **49**
5.1 Kết luận 49
5.2 Hướng phát triển 50

Tài liệu tham khảo **51**

Danh sách hình

2.1 Kiến trúc mạng U-Net [31] 5
2.2 Kiến trúc mạng FCN [26] 8
2.3 Thực hiện các thuật toán trên tập dữ liệu Gleason Challenge 2019 [[30],[19]] (Majority Voting Algorithm [34] và Staple Algorithm [34]) 14
2.4 Mô hình PSPNet [41] 15
2.5 Kiến trúc mạng ResNet [14] 15

3.1 Kiến trúc của Convolutional Block Attention Module[38] . 21
3.2 **Sơ đồ từng phần của CBAM[38]**. Như được minh họa, mô-đun channel attention sử dụng cả đầu ra gộp tối đa và đầu ra gộp trung bình với mạng chia sẻ, mô-đun spatial attention sử dụng hai đầu ra tương tự được gộp lại dọc theo trục channel và đi qua lớp tích chập. 22
3.3 Kiến trúc Atrous [3] 24
3.4 Kiến trúc Atrous Spatial Pyramid[4] 25
3.5 Các kênh đầu vào được tách riêng biệt, và mỗi kênh được thực hiện phép tích chập với bộ lọc không gian 3x3 (depth-wise). Sau đó các kênh đã được tách được nối lại, và sau đó thực hiện phép tích chập 1x1 theo điểm (pointwise) để tạo ra số lượng kênh đầu ra mong muốn.[6] 25
3.6 Cách hoạt động của tích chập theo chiều sâu 26
3.7 Cách hoạt động tích chập điểm 28
3.8 Kiến trúc Atrous Separable Spatial Pyramid Pooling(ASSPP) 29
3.9 Kiến trúc Global Attention Upsample (GAU)[22] 30
3.10 Kiến trúc Atrous Separable Spatial Pyramid Pooling (AS-SPPM) 31
3.11 Kiến trúc mạng đề xuất 34

3.12 Minh họa aux trong mạng resnet101. Với aux loss được thêm vào ngay block res4 36

4.1 Các cấp độ Gleason: (A) lành tính; (B) tuyến phôi hình thành tốt (Gleason cấp độ 3); (C) tuyến phôi hình thành kém (Gleason cấp độ 4); (D) cribriform (Gleason cấp độ 4); (E) tuyến phôi nhỏ kết hợp (Gleason cấp độ 4); (F) tuyến phôi lớn kết hợp (Gleason cấp độ 4); (G) ung thư trong ống dẫn (Gleason cấp độ 4); (H) tuyến phôi hình thành kém và tế bào đơn lẻ (Gleason cấp độ 4 và 5) [23] 38
4.2 hình ảnh microarray mô (TMA) (tissue micro-array (TMA) image) 40
4.3 Độ đo mean IoU [17] 41
4.4 Các kỹ thuật tăng cường dữ liệu. 42
4.5 Biểu đồ thể hiện giá trị Loss trong quá trình huấn luyện mô hình 45
4.6 Biểu đồ thể hiện giá trị mean IoU [17] trong quá trình huấn luyện mô hình 46
4.7 So sánh kết quả thực nghiệm của mô hình đề xuất với mô hình gốc 47

Danh sách bảng

2.1 Xếp hạng của thử thách phân đoạn EM [28] (6/3/2015) . . 6
2.2 Kết quả phân đoạn từ thử thách theo dõi tế bào của hội nghị ISBI 2015 [8]. 6
2.3 Kết quả phân đoạn của các mô hình cơ sở trên tập dữ liệu Kvarsir Intrument [18]. 7
2.4 Xếp hạng trên bộ thử nghiệm PASCAL VOC 2011 và 2012 [[10],[9]] 9
2.5 Kết quả trên bộ dữ liệu RBG-D (NYUDv2) [32] 9
2.6 Kết quả trên bộ dữ liệu SIFT Flow 9
2.7 Kết quả trên PASCAL-Contex [29] với 59 lớp nhiệm vụ . . 10
2.8 Xếp hạng trên bộ dữ liệu ADE20K [42] từ ImageNet Scene Parsing Challenge 2016 16
2.9 Kết quả đánh giá theo mIoU trong tập thử nghiệm PASCAL VOC 2012 [9]. Các phương pháp được đào tạo trước trên MS-COCO [25] được đánh dấu bằng "$\neq$". 17
2.10 Kết quả trên bộ dữ liệu thử nghiệm CityScape [7]. Phương pháp được đào tạo trên cả dữ liệu mịn và thô được đánh dấu với "$\neq$" 18

3.1 Bảng so sánh thông số cài đặt α trên ResNet50 [14]. Qua đó thấy được α có giá trị là 0.4 tối ưu nhất. 35

4.1 Kết quả thực nghiệm của mô hình PSPNet[41] (độ đo mean IoU [17] và Accurancy) 44
4.2 Bảng so sánh kết của của mô hình trên tập dữ liệu Gleason [30] [19] 48

Tóm tắt

Phân đoạn hình ảnh bệnh lý ung thư tuyến tiền liệt trong hình ảnh từ việc kiểm tra siêu vi kính của mẫu mô sinh thiết đã được nhuộm màu đóng vai trò quan trọng trong sự phát triển của lĩnh vực chuẩn đoán ung thư tuyến tiền liệt nói riêng cũng như các loại ung thư nói chung. Trong quá trình chuẩn đoán ung thư, muốn xác định được một người bệnh có mắc ung thư, cần xác định sự bất thường từ hình ảnh của mẫu mô sinh thiết, mô của người bị ung thư sẽ có sự khác biệt so với các mô bình thường. Do đó, tác vụ phân đoạn hình ảnh của mẫu mô được sinh ra để giải quyết thách thức nêu trên, cung cấp thông tin chính xác về hình ảnh mô cơ quan là bất thường hay không. Thông tin này có thể được sử dụng để cải thiện độ chính xác và hiệu quả của việc chuẩn đoán bệnh ung thư. Vì vậy, sự thành công của việc phân đoạn hình ảnh bệnh lý ung thư tuyến tiền liệt trong hình ảnh của mẫu mô sinh thiết là một trong những yếu tố quan trọng đối với sự tiến bộ của việc chuẩn đoán ung thư.

Trong quá trình phát triển của lĩnh vực học sâu, có rất nhiều mô hình áp dụng cho bài toán phân đoạn ảnh đã xuất hiện và mang lại độ chính xác cao. Một số mô hình tiêu biểu liên quan đến bài toán phân đoạn hình ảnh như: **UNet** [**31**], **FCN** [**26**], **PSPNet** [**41**], v.v ... Trong số những mô hình kể trên, nhóm nhận thấy mô hình **PSPNet** [**41**] có khả năng ứng dụng vào bài toán phân đoạn hình ảnh bệnh lý ung thư tế bào tuyến tiền liệt và đồng thời có tiềm năng để cải tiến. Do đó nhóm đã quyết định lựa chọn mô hình này để nghiên cứu và thực hiện cải tiến.

Trong khóa luận này, nhóm đã nghiên cứu và đề xuất một mô hình mới dựa trên mô hình gốc là PSPNet, có thực hiện thay đổi với backbone mới là **Resnext50-32x4d** [**39**] đồng thời cải tiến thêm với mô-đun **Convolutional Block Attention Module** [**38**] và sử dụng thêm **Atrous Separable convolution Spatial Pyramid Pooling** [**4**] là một mô-đun tích chập, áp dụng đồng thời với mô-đun giải mã **Global Attention Upsample** [**22**], bên cạnh đó cũng thực hiện các tác vụ xử lý hình ảnh và

tăng cường dữ liệu nhằm tăng độ chính xác trong quá trình đánh giá.

Nhóm đã tiến hành huấn luyện và thực nghiệm mô hình đề xuất trên tập dữ liệu Gleason là: **Gleason Challenge 2019** [[**30**], [**19**]]. Phần thực nghiệm sẽ được đánh giá dựa trên độ đo mIoU, kết quả phân đoạn đã được cải thiện đáng kể so với mô hình **PSPNet** [**41**] gốc cho thấy sự hiệu quả của phương pháp cải tiến.

Chương 1

GIỚI THIỆU ĐỀ TÀI

1.1 Đặt vấn đề

Ung thư tuyến tiền liệt **(PCa)** [**36**] là căn bệnh thứ sáu phổ biến nhất và gây tử vong thứ hai nhiều nhất ở nam giới trên toàn thế giới. Có nhiều kỹ thuật khác nhau để phát hiện và đánh giá PCa. Tuy nhiên, việc kiểm tra siêu vi kính của mẫu mô sinh thiết đã được nhuộm màu bởi các bác sĩ bệnh lý là phương pháp chính xác nhất. Dựa trên các mẫu histological có thể quan sát được, từng vùng của mô được gán một Gleason grade từ 1 đến 5. Điểm Gleason cuối cùng được báo cáo là tổng của hai mẫu rõ nét nhất và mẫu thứ hai rõ nét nhất; ví dụ, một mẫu mô có mẫu rõ nét nhất là Gleason grade 4 và mẫu thứ hai rõ nét nhất là Gleason grade 3 sẽ có điểm Gleason là 4+3.

Việc đánh giá Gleason của ung thư tuyến tiền liệt thường được thực hiện thông qua kiểm tra bằng mắt (bằng kính hiển vi) mẫu mô tuyến tiền liệt bởi các bác sĩ bệnh lý chuyên gia. Tuy nhiên, đây là một nhiệm vụ tốn thời gian và có sự biến động lớn giữa các bác sĩ, vì vậy các phương pháp tự động hỗ trợ máy tính có tiềm năng cải thiện tốc độ, độ chính xác và khả năng tái lập của kết quả.

1.2 Mục tiêu nghiên cứu

Với sự phát triển của lĩnh vực học sâu, rất nhiều mô hình học sâu áp dụng cho bài toán phân đoạn ảnh đã suất hiện. Trong vài năm trở lại đây, các mô hình đã được áp dụng nhiều hơn trong lĩnh vực y học để thực hiện các tác vụ phân đoạn các tế bào bất thường trong hình ảnh bệnh lý. Mặc dù độ chính xác của các mô hình ngày càng được cải thiện nhưng việc đưa vào thực tiễn vẫn còn gặp nhiều thách thức to lớn.

Với đề tài "Xác định tế bào ung thư tuyến tiền liệt từ hình ảnh bệnh lý" đã đặt mục tiêu quan trọng là phát triển và ứng dụng các giải pháp tự động hóa trong việc đưa ra dự đoán về các phân vùng bị ung thư từ hình ảnh bệnh lý. Thách thức phản ảnh sự cần thiết của việc sử dụng trí tuệ nhân tạo để cải thiện quá trình chuẩn đoán và dự đoán **(PCa)** [**36**], một căn bệnh quan trọng và phổ biến ảnh hưởng đến nam giới trên toàn thế giới.

Mục tiêu cuối cùng là giúp cải thiện tốc độ, độ chính xác và tính đáng tin cậy của phân vùng ung thư, từ đó tạo điều kiện thuận lợi cho các bác sĩ trong việc đưa ra quyết định điều trị và dự đoán tiến triển của bệnh cho bệnh nhân.

1.3 Phạm vi đề tài

Nội dung nghiên cứu chính của đề tài là tập trung nghiên cứu mô hình học sâu đã đạt kết quả khá tốt gần đây là mô hình PSPNet. Đầu tiên, nhóm sẽ tìm hiểu lý thuyết, thực nghiệm, phân tích các ưu điểm và khuyết điểm của mô hình. Từ đó nhóm sẽ nghiên cứu và đề xuất một số phương pháp để cải tiến mô hình.

Quá trình trên thực hiện dựa trên tập dữ liệu Gleason là: **Gleason 2019 Challenge** [[**30**],[**19**]].

1.4 Đóng góp của đề tài

Các đóng góp cụ thể của đề tài:

- Nghiên cứu tổng quan bài toán xác định ung thư tuyến tiền liệt từ hình ảnh bệnh lý sử dụng mạng phân tích kim tự tháp nói riêng cũng như các bài toán liên quan đến hình ảnh bệnh lý ung thư nói chung và một số mô hình tiêu biểu khác trong lĩnh vực phân đoạn hình ảnh.

- Nghiên cứu cơ chế hoạt động, điểm mạnh và điểm yếu của mô hình **PSPNet** [**41**] và áp dụng vào bài toán này.

- Đề xuất mô hình mới và một số phương pháp cải tiến khác nhằm cải thiện hiệu quả cách nhận diện ung thư tuyến tiền liệt.

- Thực nghiệm, so sánh kết quả của những phương pháp cải tiến trên tập dữ liệu đã nêu và chọn ra phương pháp tối ưu nhất.
- So sánh kết quả tối ưu nhất chạy thực nghiệm đạt được với mô hình đã được công bố.

1.5 Cấu trúc khóa luận

Khóa luận được chia làm 5 chương, nội dung của mỗi chương như sau:

- Chương 1 - Giới thiệu đề tài: Đặt vấn đề, trình bày mục tiêu nghiên cứu, phạm vi đề tài và đóng góp của đề tài.
- Chương 2 - Các công trình liên quan: Giới thiệu các phương pháp, công trình nghiên cứu liên quan đến bài toán phân đoạn ảnh y khoa. Phân tích, nghiên cứu và chọn ra các phương pháp phù hợp để nghiên cứu và cải tiến dựa trên mô hình gốc là **PSPNet** [**41**].
- Chương 3 - Phương pháp đề xuất: Trình bày các thay đổi để cải tiến mô hình **PSPNet** [**41**] nhóm đã thực hiện.
- Chương 4 - Kết quả thực nghiệm: Giới thiệu về môi trường thực nghiệm, các cơ sở dữ liệu liên quan, độ đo sử dụng để đánh giá, chi tiết về quy trình thực nghiệm, trình bày và nhận xét kết quả thu được.
- Chương 5 - Kết luận và hướng phát triển: Tổng kết các kết quả thu được và đóng góp của khóa luận, từ đó đề xuất các hướng phát triển cho phương pháp đề xuất của nhóm trong tương lai.

Chương 2

CÁC CÔNG TRÌNH LIÊN QUAN

Trong chương này, nhóm sẽ giới thiệu ba mô hình là **U-Net** [**31**] , **FCN** [**26**], và **PSPNet** [**41**]. Những mô hình này được sử dụng khá phổ biến trong bài toán phân đoạn ảnh nói chung cũng như phân đoạn ảnh y khoa nói riêng. Bên cạnh đó là một số thuật toán dùng cho quá trình xử lý hình ảnh như Majority Voting ... Đối với các mô hình thì đầu tiên phải kể đến **UNet** [**31**], là mô hình sơ khai nhất. Bên cạnh đó có **FCN** [**26**], là một kiến trúc được sử dụng chủ yếu để phân đoạn theo ngữ nghĩa. Cuối cùng, mô hình PSPNet cũng là một mô hình được đánh giá tương đối mạnh với các tác vụ phân đoạn hình ảnh, **PSPNet** [**41**] được phát triển với các điểm mới và kết hợp với các kiến trúc từ các mô hình nổi bật trước đó. Đối với từng mô hình được đề cập, nhóm sẽ trình bày về kiến trúc, phân tích điểm mạnh và điểm yếu, cũng như minh chứng một số kết quả thực nghiệm. Qua đó, nhóm sẽ đưa ra bàn luận và chọn mô hình thích hợp để thực hiện nghiên cứu.

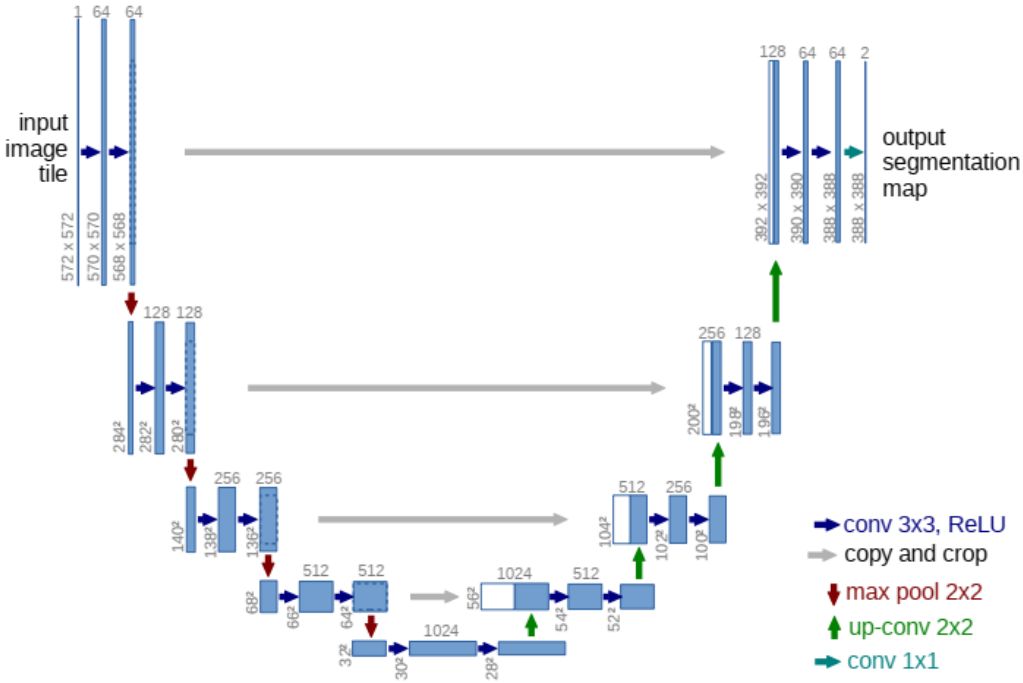
2.1 U-Net [31]

U-Net [**31**] là một mô hình học sâu phổ biến trong lĩnh vực phân đoạn ảnh y khoa dựa vào hiệu suất vượt trội vào thời điểm đầu ra mắt. Nó được công bố vào năm 2015 bởi Ronneberger và các cộng sự. Từ đó, nó được ứng dụng rộng rãi trong các tác vụ phân đoạn trong ảnh y khoa như: phân đoạn tế bào, phân đoạn khối u, phân đoạn mạch võng mạc, v.v ...

2.1.1 Kiến trúc mô hình

Kiến trúc của mô hình **U-Net** [**31**] được minh họa ở Hình 2.1. Bao gồm hai phần, phần thu hẹp (bên trái) và phần mở rộng (bên phải). Phần thu hẹp tuân theo kiến trúc điển hình của mạng nơ-ron tích chập. Mỗi

bước của nó bao gồm hai lớp tích chập 3×3 (không sử dụng padding), theo sau mỗi lớp là một hàm kích hoạt ReLU, một lớp max pooling 2×2 với bước trượt là 2 để giảm mẫu. Số kệnh đặc trưng tăng gấp đôi ở mỗi bước giảm mẫu. Mặt khác, mỗi bước của phần mở rộng lại bao gồm một lớp tăng mẫu cho bản đồ đặc trưng, mỗi lớp được theo sau bởi: một lớp tích chập tăng (up-convolution) 2×2 để làm giảm một nửa số kênh đặc trưng, một phép nối với một bản đồ đặc trưng tương ứng được cắt từ phần thu hẹp, và hai lớp tích chập 3×3 với hàm kích hoạt ReLU. Kỹ thuật cắt ảnh cần được sử dụng bởi sự mất mát của các điểm ảnh ở viền sau mỗi lần tích chập. Cuối cùng, một lớp tích chập 1×1 được sử dụng để ánh xạ từng kệnh đặc trưng trong 64 kênh thành số lớp mong muốn. Kiến trúc mạng **U-Net** [**31**] gồm tổng cộng 23 lớp tích chập.



Hình 2.1: Kiến trúc mạng U-Net [31]

Mỗi hộp xanh dương tương ứng với một bản đồ đặc trưng đa kênh, số lượng kênh được ghi chú trên từng hộp, kích thước ảnh được ghi bên trái phía dưới của hộp. Hộp trắng đại diện cho bản đồ đặc trưng sao chép. Các mũi tên thể hiện các toán tử khác nhau đã được nêu ở phần chú thích.

2.1.2 Kết quả thực nghiệm

Trong bài báo gốc của **U-Net [31]**, nhóm tác giả đã thực nghiệm trên ba tập dữ liệu ảnh y sinh khác nhau. Tác vụ đầu tiên là phân đoạn các cấu trúc nơ-ron trong các bản ghi bằng kính hiển vi điện tử. Tập dữ liệu được sử dụng thuộc thử thách phân đoạn EM của hội nghị ISBI 2012. Họ sử dụng ba độ lỗi sau để đánh giá mô hình: "Warping error", "Rand error" và "Pixel error". Kết quả của mô hình **U-Net [31]** trên tập dữ liệu EM được thể hiện ở Bảng 2.1, sắp xếp theo cột "Warping Error". Có thể thấy **U-Net [31]** độ lỗi "Warping Error" thấp nhất.

Bảng 2.1: Xếp hạng của thử thách phân đoạn EM [28] (6/3/2015)

Rank	Group	Warping Error	Rand Error	Pixel Error
	human values	0.000005	0.0021	0.0010
1.	U-Net [31]	**0.000353**	0.0382	0.0611
2.	DIVE-SCI	0.000355	0.0305	0.0584
3.	IDSIA	0.000420	0.0504	0.0613
4.	DIVE	0.000430	0.0545	**0.0582**
...				
10.	IDSIA-SCI	0.000653	**0.0189**	0.1027

Hai tập dữ liệu còn lại thuộc tác vụ phân đoạn tế bào, gồm: "PhC-U373" và "DIC-HeLa". Tác vụ này thuộc một phần của thử thách theo dõi tế bào của hội nghị ISBI vào năm 2014 và 2015 [[27], [8]]. Độ đo được sử dụng để đánh giá trên hai tập dữ liệu này là IoU, kết quả cụ thể được thể hiện ở Bảng 2.2. Ở cả hai tập dữ liệu này, kết quả của mô hình **U-Net [31]** hoàn toàn vượt trội so với mô hình tốt thứ hai.

Bảng 2.2: Kết quả phân đoạn từ thử thách theo dõi tế bào của hội nghị ISBI 2015 [8].

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best (2015)	0.83	0.46
U-Net (2015)	**0.9203**	0.7756

Ngoài ra, khi công bố tập dữ liệu dụng cụ phẫu thuật **Kvasir Instrument [18]**, tác giả đã sử dụng **U-Net [31]** như một trong những mô hình

cơ sở để thử nghiệm hiệu quả phân đoạn trên tập dữ liệu. Mục đích là để các nghiên cứu trong tương lai có thể thực hiện so sánh kết quả phân đoạn. Kết quả thực nghiệm của tác giả trên tập dữ liệu **Kvasir Instrument** được thể hiện ở Bảng 2.3. Độ đo được sử dụng gồm: Dice Similarity Coefficient (DSC), Intersection over Union (IoU) [17], precision, recall, accuracy, F2-score và Frames Per Second (FPS). Từ bảng kết quả, **U-Net [31]** hoàn toàn vượt trội so với mô hình cơ sở còn lại trên mọi độ đo.

Bảng 2.3: Kết quả phân đoạn của các mô hình cơ sở trên tập dữ liệu Kvarsir Intrument [18].

Method	IoU	DSC	F2-score	Precision	Recall	Acc
U-Net [31]	**0.8578**	**0.9158**	**0.9320**	**0.8998**	**0.9487**	**0.9864**
DoubleUnet	0.8430	0.9038	0.9147	0.8966	0.9275	0.9838

2.2 FCN - Fully Convolutional Networks [26]

Mạng tích chập hoàn toàn là một lớp mô hình phong phú giải quyết nhiều nhiệm vụ theo pixel. **FCN [26]** dành cho phân đoạn theo ngữ nghĩa cải thiện đáng kể độ chính xác bằng cách chuyển các trọng số của bộ phân loại được huấn luyện trước, kết hợp các biểu diễn lớp khác nhau và học từ đầu đến cuối trên toàn bộ hình ảnh. Hoạt động từ đầu đến cuối, từ điểm ảnh đến điểm ảnh đồng thời đơn giản hóa và tăng tốc độ học tập và suy luận.

2.2.1 Kiến trúc mô hình

Mạng tích chập hoàn toàn, hay **FCN [26]**, là một kiến trúc được sử dụng chủ yếu để phân đoạn theo ngữ nghĩa. Chúng chỉ sử dụng các lớp được kết nối cục bộ, chẳng hạn như tích chập, gộp và giải mã. Tránh sử dụng các lớp dày đặc có nghĩa là ít tham số hơn làm cho mạng được đào tạo nhanh hơn. Điều đó cũng có nghĩa là **FCN [26]** có thể hoạt động với các kích thước hình ảnh khác nhau do tất cả các kết nối đều cục bộ.

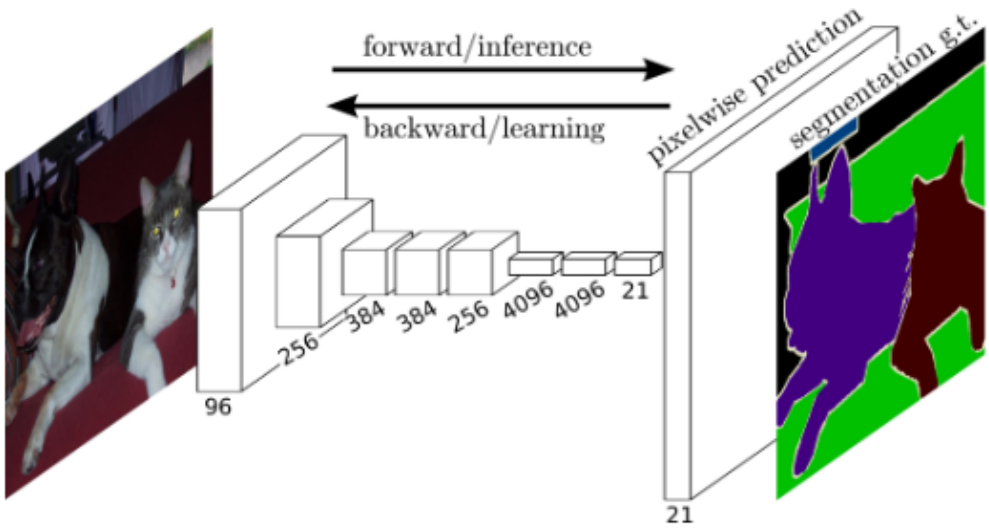
Kiến trúc gồm 2 path có tác dụng khác nhau đó là:

- Downsampling path: Trích suất đặc trưng ngữ cảnh và nội dung của ảnh. Downsampling path là kiến trúc thu hẹp kích thước lớp.

Đó chính là lớp CNN ở vị trí đầu tiên, thông thường là những kiến trúc CNN cơ bản như **AlexNet** [**20**], **VGG16** [**33**], **VGG19** [**33**], **ResNet** [**14**] đã được lược bớt kết nối đầy đủ. Đầu ra của Downsampling path là bản đồ đặc trưng (Feature Map) mang giá trị thông tin về vị trí và cường độ của các pixels. Tích chập kích thước 1 x 1 được áp dụng ngay sau khối CNN với số lượng bộ lọc bằng với số lượng nhãn ở đầu ra.

- Upsampling path: Khôi phục thông tin không gian của ảnh theo các phân đoạn dự báo. Chúng ta áp dụng một lớp giải chập để giải chập thông tin thành giản đồ phân đoạn.

FCN [**26**] cũng sử dụng các kết nối tắt để khôi phục thông tin không gian chi tiết bị mất trong phần giảm mẫu.



Hình 2.2: Kiến trúc mạng FCN [26]

2.2.2 Kết quả thực nghiệm

Trong bài báo gốc của **FCN** [**26**], nhóm tác giả đã thực nghiệm trên các tập dữ liệu ảnh khác nhau. Đầu tiên với **PASCAL VOC**, cho biết hiệu suất FCN-8 trên bộ thử nghiệm **PASCAL VOC 2011 và 2012** [[**10**],[**9**]], và so sánh nó với phiên bản tốt nhất trước đó. Kết quả của mô hình FCN-8 trên tập dữ liệu **PASCAl VOC** được thể hiện ở Bảng 2.4, mô hình đạt được kết quả tốt nhất trên mean IU.

Bảng 2.4: Xếp hạng trên bộ thử nghiệm PASCAL VOC 2011 và 2012 [[10],[9]]

	meanIU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN	47.9	-	-
SDS	52.6	51.6	$\sim$ 50 s
FCN-8s	**67.5**	**67.2**	$\sim$ **100 ms**

Tiếp theo là trên bộ liệu **NYUDv2 [32]** là bộ dữ liệu RGB-D được thu thập bằng cách sử dụng Microsoft Kinect. Có 1.449 hình ảnh RGB-D, với pixelwise các nhãn đã được kết hợp thành nhiệm vụ phân đoạn ngữ nghĩa trên 40 lớp. Kết quả được thể hiện ở Bảng 2.5 với mô hình FCN-32s RGB-HHA vượt trội trên các độ đo pixel accuracy, mean accuracy, mean IU, f.w.IU

Bảng 2.5: Kết quả trên bộ dữ liệu RBG-D (NYUDv2) [32]

	pixel acc.	mean acc.	mean IU	f. w. acc.
Gupta et al. [12]	60.3	-	28.6	47.0
FCN-32s RGB	61.8	44.7	31.6	46.0
FCN-32s RGB-D	62.1	44.8	31.7	46.3
FCN-32s HHA	58.3	35.7	25.2	41.7
FCN-32s RGB-HHA	**65.3**	**44.0**	**33.3**	**48.6**

SIFT Flow là tập dữ liệu gồm 2.688 hình ảnh có nhãn pixel cho 33 lớp ngữ nghĩa (“cây cầu”, “núi”, “mặt trời”), như cũng như ba lớp hình học (“ngang”, “dọc” và "bầu trời"). Kết quả với các độ đo tương tự như bộ dữ liệu **NYUDv2 [32]** và có thêm một độ đo là geom. accuracy, mô hình **FCN-8s** có sự vượt trội trên các độ đo, được thể hiện ở Bảng 2.6

Bảng 2.6: Kết quả trên bộ dữ liệu SIFT Flow

	pixel acc.	mean acc.	mean IU	f. w. acc.	geom. acc.
Liu et al.	76.7	-	-	-	-
Tighe et al transfer	-	-	-	-	90.8
Tighe et al SVM	75.6	41.1	-	-	-
Tighe et al SVM+MRF	78.6	39.2	-	-	-
Farabet et al natural	72.3	50.8	-	-	-
Farabet et al balanced	78.5	29.6	-	-	
Pinheiro et al	77.7	29.8	-	-	-
FCN-8s	**85.9**	**53.9**	41.2	77.2	**94.6**

PASCAL-Context [29] cung cấp chú thích toàn cảnh của PASCAL VOC 2010. Trong Bảng 2.7, tác giả đã so sánh với kết quả tốt nhất trước đó cho nhiệm vụ này. FCN-8 đạt 39,1 mean IU cho một cải thiện tương đối lớn hơn 10 %.

Bảng 2.7: Kết quả trên PASCAL-Contex [29] với 59 lớp nhiệm vụ

	pixel acc.	mean acc.	mean IU	f. w. acc.
O2P	-	-	18.1	-
CFM	-	-	34.4	-
FCN-32s	65.5	49.1	36.7	50.9
FCN-16s	66.9	51.3	38.4	52.3
FCN-8s	**67.5**	**52.3**	**39.1**	**53.0**

Mô hình **FCN [26]** được thực nghiệm trên nhiều bộ dữ liệu đa dạng nhưng vẫn đạt được hiệu suất cao vượt trội và ổn định so với các mô hình trước đó, với các phiên bản của mô hình FCN là FCN-32s, FCN-16s và FCN-8s.

2.3 Các thuật toán liên quan

Các Ký hiệu:

- g: Tập hợp tất cả các điểm ảnh có thể được gán các lớp Gleason trên một hình ảnh, được định nghĩa là g = 0, 3, 4, 5.
- h: Chiều cao của hình ảnh đầu vào.
- w: Chiều rộng của hình ảnh đầu vào.
- p: Tập hợp tất cả các giá trị có thể của một điểm ảnh trong một kênh hình ảnh RGB, được định nghĩa là p = 0, 1, . . . , 255.
- $p^{3\times h\times w}$: Không gian của tất cả các hình ảnh đầu vào với các kênh RGB.
- $g^{h\times w}$: Không gian của tất cả các kết quả phân đoạn có thể có.

- f: Hàm f: $f : p^{3\times h\times w} \rightarrow g^{h\times w}$ được định nghĩa là hàm nhận một hình ảnh đầu vào và trả về một kết quả phân đoạn tương ứng bằng các lớp Gleason 0, 3, 4 và 5.

- $\{Map_i\}_{i=1}^{6}$: Tập chứa các phiên bản của các hàm được đề cập, trong đó Map_i nhận một hình ảnh đầu vào từ tập dữ liệu và trả về kết quả phân đoạn được cung cấp trong bản đồ thứ i.

Mục Tiêu: Định nghĩa một hàm f nhận hình ảnh RGB làm đầu vào (với không gian là $p^{3\times h\times w}$), xử lý và đưa ra kết quả phân đoạn ngữ nghĩa (với không gian là $g^{h\times w}$) sử dụng các lớp Gleason cụ thể (0, 3, 4 và 5). Hàm f được thiết kế để cung cấp kết quả phân đoạn chính xác và ý nghĩa cho hình ảnh lõi TMA trên tập dữ liệu **Gleason Challenge 2019** [[**30**],[**19**]].

2.3.1 Majority Voting Algorithm [34]

Thuật toán **Majority Voting** [**34**] đề xuất một hàm mục tiêu như là sự phân loại lỗi giữa sáu bản đồ và đặt ra vấn đề về việc tìm một hàm f để giảm thiểu giá trị của hàm mục tiêu [21]. Hàm mục tiêu được định nghĩa như sau:

$$\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I(f(x)_{ij} \neq Map_k(x)_{ij} \quad \textbf{(1)}$$

Sau khi biến đổi hàm mục tiêu để tối thiểu hóa lỗi phân loại, tương đương với việc tối đa hóa biểu thức:

$$\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I(f(x)_{ij} = Map_k(x)_{ij} \quad \textbf{(2)}$$

Hàm mục tiêu được tối đa hóa khi chọn phần lớn bỏ phiếu đa số của tất cả sáu bản đồ cho mỗi điểm ảnh của một hình ảnh. Do đó, thuật toán bỏ phiếu đa số định nghĩa *MajVote(x)*: $p^{3\times h\times w} \rightarrow g^{h\times w}$

$$MajVote(x) := \arg\max_f \sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{6} I(f(x)_{ij} = Map_k(x)_{ij} \quad \textbf{(3)}$$

Mặc dù đơn giản và thành công trong việc tìm ra một giải pháp tối ưu toàn diện cho hàm mục tiêu được đề xuất, thuật toán bỏ phiếu đa số vẫn có nhược điểm như sau:

- Thứ nhất: Hàm mục tiêu trong (2) cho thấy rằng lớp của mỗi điểm ảnh chỉ phụ thuộc vào điểm ảnh tương ứng trong các bản đồ khác nhau. Điều này cho thấy rằng phương trình bỏ phiếu đa số trong (3) không thể đảm bảo rằng các điểm ảnh khác nhau trong mỗi vùng ngữ nghĩa chia sẻ cùng một lớp.

- Thứ hai: Thuật toán bỏ phiếu đa số cho thấy rằng các phân đoạn chứa các lớp có tần suất thấp nên bị loại bỏ khỏi đại diện cuối cùng, và chỉ các phân đoạn mà được sự đồng ý cao mới nên được xem xét. Do đó, có sự thiên vị hoặc ưu tiên trong việc chọn các lớp có tần suất cao có thể không phù hợp cho các vấn đề như đánh giá ung thư tuyến tiền liệt trong hình ảnh TMA.

- Thứ ba: Sự ưu thế của thuật toán bỏ phiếu đa số giảm khi áp dụng vào một tập dữ liệu không cân bằng, điều này xảy ra với tập dữ liệu **Gleason Challenge 2019** [[**30**],[**19**]]. Hậu quả là nhiều phân đoạn bị loại bỏ khỏi định danh thực thể cuối cùng, làm cho lớp bị loại bỏ nhiều phân đoạn trở nên hiếm hơn so với tập hợp ban đầu của các phân đoạn. Cuối cùng, thuật toán này giảm biến động của các mẫu lớp trong tập dữ liệu, dẫn đến việc huấn luyện kém chất lượng cho bất kỳ mô hình học sâu nào.

Thuật toán bỏ phiếu đa số được thực hiện trên các bản đồ phân đoạn của các hình ảnh khác nhau từ tập dữ liệu Gleason Challenge 2019 và hiển thị kết quả trong Hình 2.3. Ở cột **MajVot** liên quan đến thuật toán bỏ phiếu đa số được áp dụng vào các bản đồ phân đoạn. Thấy được các vùng có độ tương phản cao, thuật toán đã tạo ra các hình dạng không bình thường loại bỏ các thuộc tính ngữ nghĩa cần thiết cho các vùng của các lớp đó. Và có thể thấy các lớp có tần suất thấp như lớp Gleason 5 đã bị loại hoàn toàn khỏi đầu ra của thuật toán này.

2.3.2 STAPLE Algorithm [34]

Thuật toán **STAPLE** [**34**] xây dựng bản đồ thực tế cuối cùng như một thuật toán EM [21], lặp lại cập nhật phân đoạn đã xây dựng để tối đa hóa hàm **log-likelihood** dữ liệu không đầy đủ. Cụ thể, thuật toán ban đầu

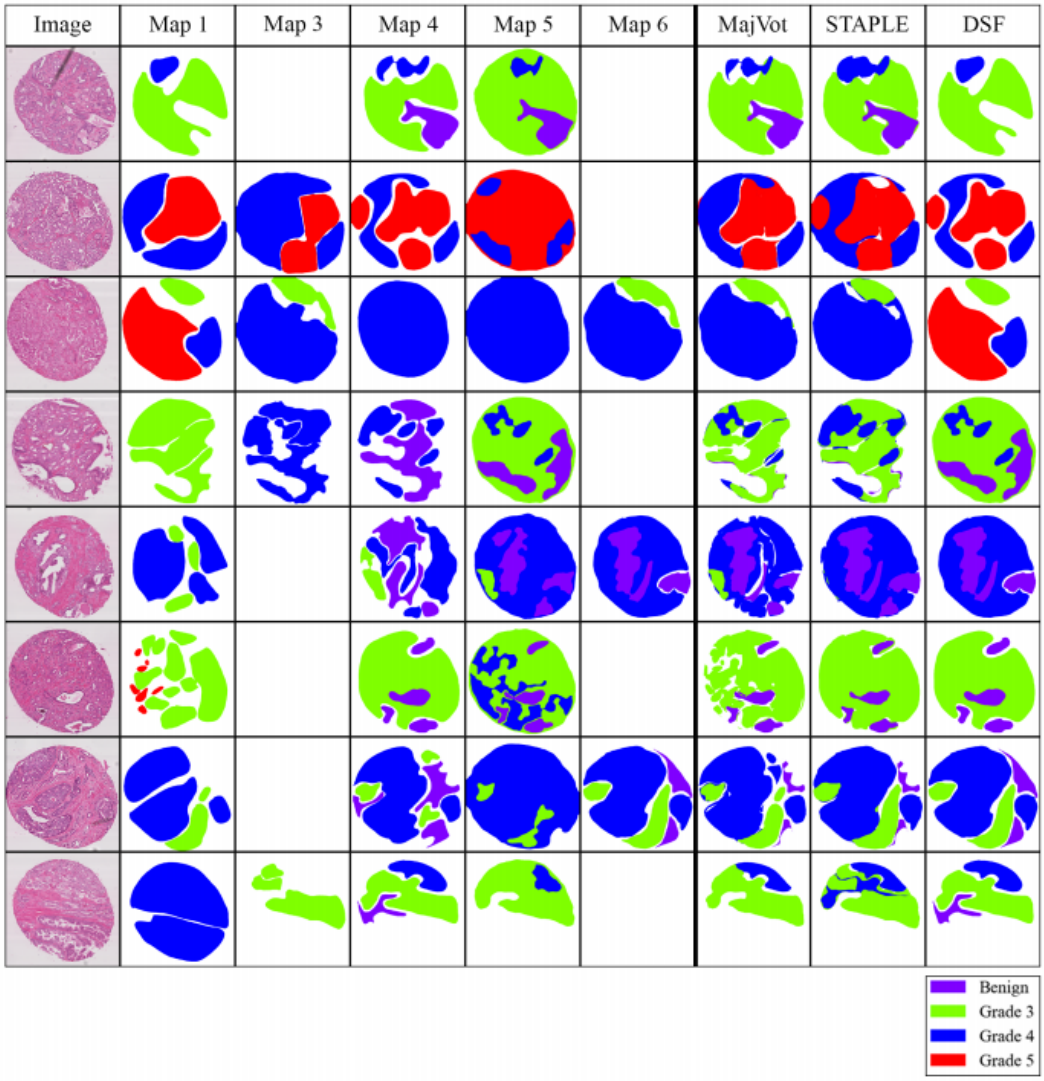
xây dựng một phân đoạn thử nghiệm bằng cách bỏ phiếu đơn giản trên mỗi điểm ảnh. Hơn nữa, thuật toán tính toán các tham số hiệu suất của mỗi chú thích so với phân đoạn thử nghiệm, là bước E, và xây dựng một phân đoạn thử nghiệm mới bằng cách sử dụng trọng số được tính toán trước đó là bước M. Quá trình này được lặp lại cho đến khi phân đoạn thử nghiệm hội tụ và phân đoạn thử nghiệm đã hội tụ được coi là nhãn thực cuối cùng.

Mặc dù được áp dụng rộng rãi trên một số tập dữ liệu khác nhau như một phương pháp tổng hợp mới, hiệu suất của thuật toán này trong nhiệm vụ phân đoạn ngữ nghĩa trên bộ dữ liệu **Gleason Challenge 2019** [[**30**],[**19**]] có hai điểm yếu:

- Thứ nhất: quy trình EM được triển khai trong thuật toán **STAPLE** [**34**] không xem xét lớp được gán cho một điểm ảnh dựa trên các lớp của các điểm ảnh xung quanh. Nói cách khác, các thuộc tính ngữ nghĩa giống nhau của một kết quả phân đoạn chấp nhận được đã được thảo luận trong phần trước không được xem xét trong kết quả của thuật toán **STAPLE**.

- Thứ hai: Thuật toán **STAPLE** [**34**] hoạt động kém trên các tập dữ liệu có sự mất cân bằng cao về các loại lớp và chứa nhiều phân đoạn mâu thuẫn, điều này đã xảy ra với tập dữ liệu **Gleason Challenge 2019** [[**30**],[**19**]]. Tương tự như thuật toán bỏ phiếu đa số, thuật toán **STAPLE** thường loại bỏ lớp có tần suất ít nhất, Gleason grade 5, và giảm đáng kể tần suất của grade 5 trong các bản đồ thực tế cuối cùng.

Hình 2.3thể hiện kết quả triển khai thuật toán **STAPLE** [**34**] trên các bản đồ phân đoạn của nhiều hình ảnh từ bộ dữ liệu **Gleason Challenge 2019** [[**30**],[**19**]]. Cột **STAPLE** đề cập đến kết quả của thuật toán STAPLE được áp dụng vào các bản đồ phân đoạn. Có thể thấy rằng ở những vùng có độ tương phản cao, đã được tạo ra các hình dạng không bình thường loại bỏ các thuộc tính ngữ nghĩa cần thiết cho các vùng của những lớp đó. Hơn nữa, có thể thấy rằng các lớp có tần suất thấp như Gleason grade 5 đã bị loại bỏ hoàn toàn trong kết quả của thuật toán này.

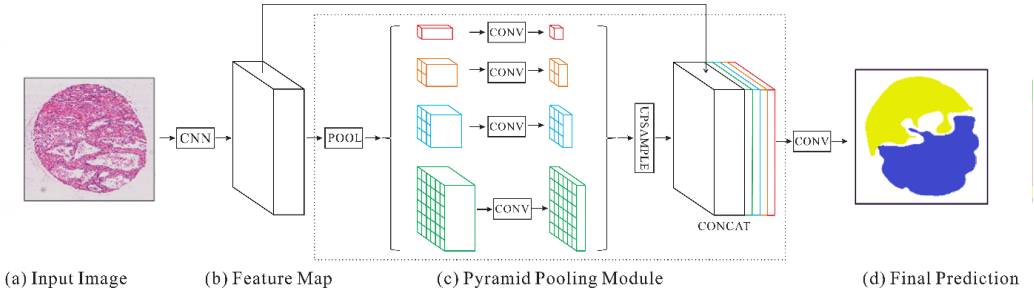


Hình 2.3: Thực hiện các thuật toán trên tập dữ liệu Gleason Challenge 2019 [[30],[19]] (Majority Voting Algorithm [34] và Staple Algorithm [34])

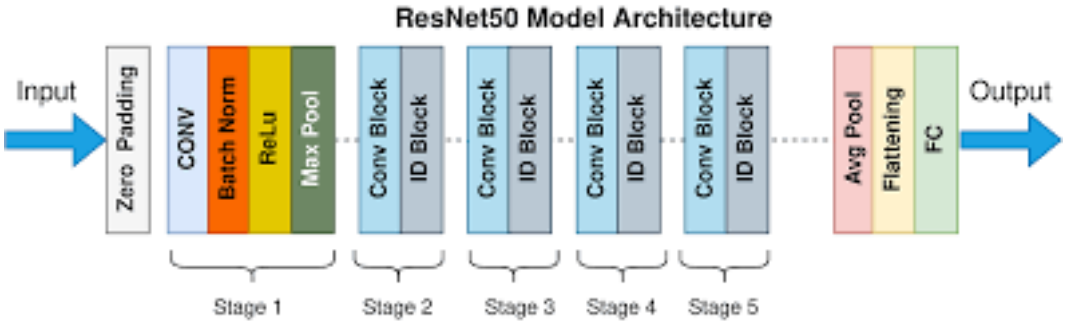
2.4 PSPNet - Pyramid Scene Parsing Network [41]

2.4.1 Kiến trúc mô hình

PSPNet [41] là một mô hình phân đoạn ngữ nghĩa sử dụng mô-đun phân tích kim tự tháp để khai thác thông tin ngữ cảnh toàn cục bằng cách tổng hợp ngữ cảnh dựa trên các khu vực khác nhau. Sử dụng mạng CNN đã được huấn luyện từ trước là **ResNet** [14] để đưa ra được bản đồ đặc trưng có kích thước bằng 1/8 so với ảnh gốc. Sau đó đi qua một mô-đun tổng hợp kim tự tháp với 4 cấp độ khác nhau sau đó được hợp nhất lại với nhau và thực hiện việc tăng mẫu để đưa về kích thước bằng với kích thước ảnh gốc.



Hình 2.4: Mô hình PSPNet [41]



Hình 2.5: Kiến trúc mạng ResNet [14]

Mục tiêu chính của mô-đun tổng hợp kim tự tháp (Pyramid Pooling Module) là nâng cao khả năng phân đoạn các vùng khác nhau trong hình ảnh với các kích thước vùng quan sát khác nhau. PPM (Pyramid Pooling Module) hoạt động bằng cách chia hình ảnh đầu vào thành các phạm vi hoặc cửa sổ con với các kích thước khác nhau, sau đó áp dụng một phép gom nhóm (pooling operation) trong từng cửa sổ con để trích xuất thông

tin cụ thể từ mỗi phạm vi. Sau đó, các thông tin từ các cửa sổ con này được tổng hợp lại để tạo ra một biểu đồ của hình ảnh ở nhiều tỷ lệ kích thước khác nhau. Điều này giúp mô hình có khả năng phân đoạn các vật thể ở các kích thước và tỷ lệ khác nhau trong hình ảnh.

2.4.2 Kết quả thực nghiệm

Phương pháp đề xuất trong bài báo gốc đã thành công trong việc phân tích cảnh và những thách thức về phân đoạn theo ngữ nghĩa. Tác giả đã đánh giá mô hình trên ba bộ dữ liệu khác nhau, bao gồm **ImageNet Scene Parsing Challenge 2016** [**42**], phân đoạn ngữ nghĩa với bộ dữ liệu **PASCAL VOC 2012** [**9**] và hiểu về cảnh cảnh đô thị trên tập dữ liệu **Cityscapes** [**7**].

Đầu tiên là chạy thực nghiệm với bộ dữ liệu **ADE20K** [**42**] từ thử thách **ImageNet Scene Parsing Challenge 2016**. Bảng 2.8 thể hiện phân tích chi tiết về **PSPNet** [**41**]được đề xuất trong bài báo gốc của tác giả và so sánh với các mô hình khác. Kết quả thu được trên tập validation với đầu vào là single-scale ngoại trừ hàng cuối cùng.**'DA'** đề cập đến dữ liệu phần tăng thêm mà tác giả đã thực hiện,**'AL'** biểu thị tổn thất phụ mà tác giả đã thêm và 'PSP' đại diện cho **PSPNet** [**41**] được đề xuất. **'MS'** có nghĩa là thử nghiệm đa quy mô được sử dụng.

Bảng 2.8: Xếp hạng trên bộ dữ liệu ADE20K [42] từ ImageNet Scene Parsing Challenge 2016

Method	Mean IoU (%)	Pixel Acc. (%)
FCN [26]	29.39	71.32
SegNet	21.64	71.00
DilatedNet	32.31	73.55
CascadeNet	34.90	74.52
ResNet50-Baseline	34.28	76.35
ResNet50+DA	35.82	77.07
ResNet50+DA+AL	37.23	78.01
ResNet50+DA+AL+PSP	**41.68**	**80.04**
ResNet269+DA+AL+PSP	43.81	80.88
ResNet269+DA+AL+PSP+MS	**44.94**	**81.69**

Kết quả thử thách **ImageNet Scene Parsing Challenge 2016** [**42**],

kết quả tốt nhất của mỗi mô hình được liệt kê. Điểm đánh giá cuối cùng là giá trị trung bình của IoU [17] và Pixel Acc. Kết quả được đánh giá trên bộ thử nghiệm cho thấy mô hình trong bài báo gốc được tác giả sử dụng đạt hạng nhất với final score là 57.21 %. Vượt qua các mô hình khác và vượt trôi so với các mô hình cũ trước đó.

PSPNet [41] cũng hoạt động tốt trên phân đoạn ngữ nghĩa. Tác giả đã thực hiện đánh giá trên tập dữ liệu phân đoạn **PASCAL VOC 2012 [9]**, chứa 20 loại đối tượng và một lớp nền.

Bảng 2.9: Kết quả đánh giá theo mIoU trong tập thử nghiệm PASCAL VOC 2012 [9]. Các phương pháp được đào tạo trước trên MS-COCO [25] được đánh dấu bằng "$\neq$".

Method	mean IoU
FCN [26]	62.2
Zoom-out	69.6
DeepLab [5]	71.6
CRF-RNN	72.0
DeconvNet	72.5
GCRF	73.2
DPN	74.1
Piecewise	75.3
PSPNet [41]	**82.6**
$CRF-RNN^{\neq}$	74.7
$BoxSup^{\neq}$	75.2
$Dilation8^{\neq}$	75.3
$DPN^{\neq}$	77.5
$Piecewise^{\neq}$	78.0
$FCRNs^{\neq}$	79.1
$LRR^{\neq}$	79.3
$DeepLab^{\neq}$ [5]	79.7
$PSPNet^{\neq}$ [41]	**85.4**

Như được hiển thị trong Bảng 2.9, **PSPNet [41]** hoạt động tốt hơn các phương pháp trước đó trên cả hai cài đặt. Được đào tạo chỉ với dữ liệu **VOC 2012 [9]**, **PSPNet [41]** đạt được độ chính xác 82,6 % và đạt được độ chính xác cao nhất trên cả 20 lớp. Khi **PSPNet [41]** được đào tạo trước với tập dữ liệu **MS-COCO [25]**, nó đạt độ chính xác 85,4 % trong đó 19 trên 20 lớp nhận được độ chính xác cao nhất. Từ các kết quả thực

nghiệm có được, có thể lập luận rằng mô hình dựa trên sự phân loại này là mạnh hơn một số phương pháp trước đó kể từ **ResNet [14]** đã được đề xuất gần đây.

Cityscapes [7] là một bộ dữ liệu về ngữ nghĩa, hiểu biết về cảnh quan đô thị. Nó chứa 5.000 hình ảnh chất lượng cao có chú thích chi tiết ở cấp độ pixel được thu thập từ 50 thành phố trong các mùa khác nhau.

Bảng 2.10: Kết quả trên bộ dữ liệu thử nghiệm CityScape [7]. Phương pháp được đào tạo trên cả dữ liệu mịn và thô được đánh dấu với "$\neq$"

Method	IoU cla.	iIoU cla.	IoU cat.	iIoU cat.
CRF-RNN	62.5	34.4	82.7	6.0
FCN [26]	65.3	41.7	85.7	70.1
SiCNN	66.3	44.9	85.0	71.2
DPN	66.8	39.1	86.0	69.1
Dililation10	67.1	42.0	86.5	71.1
LRR	69.7	48.0	88.2	74.7
DeepLab [5]	70.4	42.6	86.4	67.7
Piecewise	71.6	51.7	87.3	74.1
PSPNet [41]	**78.4**	**56.7**	**90.6**	**78.6**
$LRR^{\neq}$	71.8	47.9	88.4	73.9
$PSPNet^{\neq}$ [41]	**80.2**	**58.1**	**90.6**	**78.2**

Thống kê trong Bảng 2.10 cho thấy **PSPNet [41]** vượt trội hơn các phương pháp với ưu điểm nổi bật. Được đào tạo với tập thử nghiệm **CityScape [7]**, **PSPNet [41]** đạt độ chính xác 78,4 % và tất cả 20 lớp nhận được độ chính xác cao nhất. Sử dụng cả dữ liệu mịn và thô để đào tạo thì **PSPNet [41]** mang lại độ chính xác 80.2 % và đạt được độ chính xác cao nhất trên cả 20 lớp.

2.5 Tổng kết

Thông qua việc nghiên cứu kỹ lưỡng các mô hình đã nêu, các mô hình nổi bật trong phân đoạn, phân loại hình ảnh, cũng như dựa trên kết quả thực nghiệm theo các nhóm tác giả nghiên cứu. Nhóm nhận thấy mô hình PSPNet có tiềm năng để phát triển nhất, phù hợp với bài toán phân đoạn hình ảnh ung thư tế bào tuyến tiền liệt trong hình ảnh siêu vi kính trên các mẫu mô.

Lý do chính là mô hình **PSPNet** [**41**] (Pyramid Scene Parsing Network) có một số ưu điểm riêng biệt. PSPNet thường cho hiệu suất tốt hơn trên các bộ dữ liệu lớn và có độ phức tạp cao. Tích hợp các cấu trúc thông tin từ ngữ cảnh khác nhau giúp cải thiện khả năng phân đoạn. Bên cạnh đó, **PSPNet** [**41**] cũng là mô hình thường được sử dụng cho bài toán phân đoạn.

Ngoài ra, nhận thấy mặc dù có rất nhiều mô hình đã được đề xuất và đạt kết quả khá tốt cho bài toán phân đoạn hình ảnh bệnh lý nhưng vẫn chưa có nhiều thực nghiệm để kiểm chứng độ hiệu quả của những mô hình đó cho bài toán phân đoạn tế bào ung thư.

Tóm lại, **PSPNet** [**41**] là một mô hình semantic segmentation mạnh mẽ với khả năng tổng hợp ngữ cảnh từ nhiều quy mô khác nhau và kết hợp thông tin chi tiết, cho phép mô hình học được ngữ cảnh một cách toàn diện. Điều này giúp nó đạt được hiệu suất tốt trên nhiều bộ dữ liệu và bài toán phức tạp.

Từ các lý do trên, nhóm quyết định chọn mô hình **PSPNet** [**41**] là mô hình chính để nghiên cứu, cải tiến và thực nghiệm trên bài toán phân đoạn hình ảnh ung thư tế bào tuyến tiền liệt, nhằm kiểm chứng hiệu quả phân đoạn của mô hình.

Chương 3

PHƯƠNG PHÁP ĐỀ XUẤT

3.1 Đề xuất một số kiến trúc có thể được kết hợp

3.1.1 Convolutional Block Attention Module

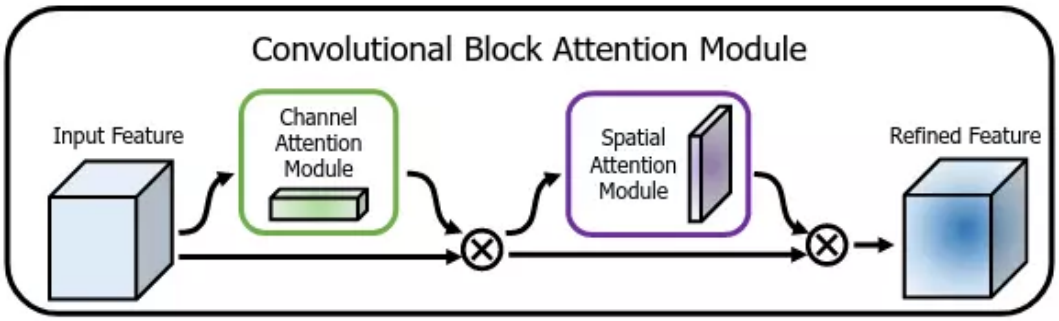
Mô-đun Convolutional Block Attention, hay CBAM [38] , là một đóng góp quan trọng trong lĩnh vực thị giác máy tính, được phát triển bởi bốn nhà nghiên cứu nổi tiếng. Đồng tác giả gồm Sanghyun Woo và In So Kweon, đại diện cho Viện Công nghệ và Khoa học tiên tiến Hàn Quốc (Korea Advanced Institute of Science and Technology - KAIST), Jongchan Park từ Lunit Inc. ở Seoul, Hàn Quốc, và Joon-Young Lee đến từ Adobe Research tại San Jose, California, Hoa Kỳ. Sự kết hợp của họ mang lại sự đa dạng về kiến thức và kinh nghiệm, tạo ra một mô-đun mạnh mẽ có khả năng tăng cường hiệu suất của các mô hình học sâu thông qua việc cải thiện quá trình trích xuất đặc trưng và chú ý.

Trong bối cảnh hiện nay, khi các phép toán tích chập trích xuất đặc trưng thông tin bằng cách kết hợp thông tin chéo kênh (cross-channel) và không gian (spatial), việc áp dụng CBAM[38] để nhấn mạnh các đặc trưng ý nghĩa theo hai chiều chính, đó là trục kênh và không gian(chanel và spatial), đại diện cho một bước tiến quan trọng. Tác giả tuần tự áp dụng các mô-đun channel và spatial, giúp mỗi nhánh của mô-đun học được **'cái gì'** và **'nơi nào'** cần được chú ý trong quá trình học tương ứng. Kết quả là, CBAM[38] cải thiện hiệu quả quá trình thông tin lưu thông trong mạng nơ-ron và đồng thời giúp quá trình học tăng cường khả năng nhấn mạnh hoặc giảm thiểu thông tin nào nên được ưu tiên trong quá trình học. Đầu vào là một features map có kích thước $F \in R^{C*H*W}$, tuần tự đi qua mô-đun channel attention thu được vector có kích thước $M_c \in R^{C*1*1}$ và mô-đun spatial attention thu được vector có kích thước $M_s \in R^{1*H*W}$.

Mô-đun CBAM[38] được minh họa ở hình3.1, minh họa bằng công thức sau đây:

$$F^{'} = M_c(F) \bigotimes F,$$
$$F^{"} = M_s(F^{'} \bigotimes) F^{'}$$

Trong đó $\bigotimes$ là phép nhân theo phần tử, giúp các giá trị chú ý được sao chép vào features map đầu vào của mô-đun. Hình3.2 mô tả quá trình tính toán của từng mô-đun trong CBAM[38]. Phần sau đây mô tả chi tiết về từng mô-đun trong CBAM.

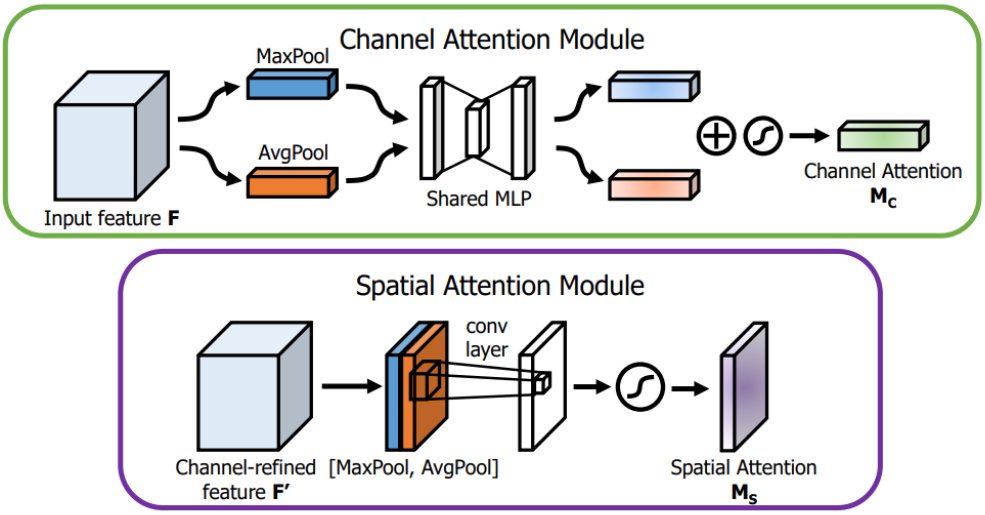


Hình 3.1: Kiến trúc của Convolutional Block Attention Module[38]

- **Channel Attention Module(CAM)**: Trong phần Squeeze của CBAM[38] sử dụng đồng thời cả GAP(Global Average Pooling)[24] và cả GMP (Global Max Pooling). Tác giả của CBAM [38] nói rằng GMP cũng thu thập thông tin quan trọng của vật thể nhưng theo một khía cạnh khác. Sau đó, các đặc trưng thu được thì GMP và GAP đều được cho đi qua chung phần Excitation tạo ra hai vector. Sau đó, hai Vector này được cộng vào với nhau và và đi qua một hàm sigmoid và trả về một bản đồ đặc trưng 1 chiều. Mục tiêu chính của CAM là tập trung vào 'cái gì' có ý nghĩa trong bản đồ đặc trưng đầu. Để tính toán CAM một cách hiệu quả, tác giả thu nhỏ tối đa kích thước không gian của bản đồ đặc trưng đầu vào và cho đi qua đồng thời cả GAP(Global Average Pooling) và cả GMP (Global Max Pooling). Đối với việc tổng hợp thông tin không gian, phương pháp lấy trung bình đã được sử dụng phổ biến cho đến nay do. Zhou et al đề xuất sử dụng nó để hiệu quả học được phạm vi của đối tượng

mục tiêu và Hu et al áp dụng nó trong mô-đun attention của họ để tính toán thống kê không gian. Bên cạnh đó tác giả nói rằng GMP cũng thu thập thông tin quan trọng của vật thể nhưng theo một khía cạnh khác về các đặc trưng đặc biệt của đối tượng để suy luận CAM mịn hơn. Quá trình tính toán được thực hiện như sau, với đầu vào là một features map $F \in R^{C*H*W}$. Sau khi đi qua GAP và GMP thì trả về 2 vector $F^c_{avg}, F^c_{max} \in R^{C*1*1}$ và được chuyển tiếp tới một mạng nơ-ron shared MLP với 1 lớp ẩn có mục đích nhằm giảm số lượng siêu tham số trong mạng. 2 vector sau đó được tính tổng bằng phép tính tổng theo phần tử và một hàm kích hoạt sigmoid. Minh họa quá trình tính toán bằng công thức sau :

$$M_c(F) = \sigma(MLP(AVGPool(F)) + MLP(MaxPool(F)))$$



Hình 3.2: **Sơ đồ từng phần của CBAM[38]**. Như được minh họa, mô-đun channel attention sử dụng cả đầu ra gộp tối đa và đầu ra gộp trung bình với mạng chia sẻ, mô-đun spatial attention sử dụng hai đầu ra tương tự được gộp lại dọc theo trục channel và đi qua lớp tích chập.

- **Spatial Attention Module(SAM)**: Khác biệt so với CAM, SAM tập trung vào 'nơi nào' là một phần thông tin quan trọng, điều này là một sự bổ sung hoàn hảo cho CAM. Cách thực hiện Spatial Attention của CBAM[38] rất đơn giản, tương tự như Channel Attention. Thay vì thực hiện GAP và GMP theo chiều channel, SAM thực hiện theo

chiều spatial của feature maps, tạo ra hai feature maps tổng hợp có số chiều channel là một. Hai feature maps này, sau đó được nối lại với nhau bằng phép concatenate và đi qua một lớp Conv với kernel size là 7, sau đó đi qua hàm Sigmoid với số filter là 1 để tạo ra bản đồ chú ý. Đáng chú ý là ta phải dùng một Conv có kernel size to để có thể tính toán được vùng thông tin rộng. Nếu ta dùng lớp Conv có kernel size là 3 thì vùng tính toán được của ta sẽ rất hẹp, giảm đi hiệu suất của bản đồ chú được tạo ra. Minh họa chi tiết bằng công thức sau :

$$M_s(F) = \sigma(f^{7*7}([AvgPool(F); MaxPool(F)]))$$

Mục đích thêm mô-đun CBAM [38] vào mô hình PSPNet [41] nhằm cải thiện khả năng học và độ chú ý đối với các đặc trưng spatial và channel của hình ảnh. CBAM[38] sử dụng giá trị chú ý được tính toán để cải thiện đầu ra của mỗi kênh theo chiều không gian và ngược lại. Điều này giúp mô hình tập trung vào các đặc trưng quan trọng, có thể giúp nâng cao khả năng phân loại và hiểu sự biểu diễn của hình ảnh đầu vào.

3.1.2 Atrous convolution Spatial Pyramid

Có nhiều biến thể mô hình được đề xuất để khai thác thông tin ngữ cảnh cho quá trình phân đoạn, bao gồm những mô hình sử dụng đầu vào đa quy mô (ví dụ: pyramid hình ảnh) hoặc những mô hình sử dụng mô hình đồ thị xác suất (như DenseCRF với thuật toán suy luận hiệu quả). Chúng tôi đề xuất về các mô hình sử dụng Spatial Pyramid Pooling [4] để kết hợp với mô-đun pyramid pooling của PSPNet[41] nhằm mở rộng vùng nhận thức (receptive field).

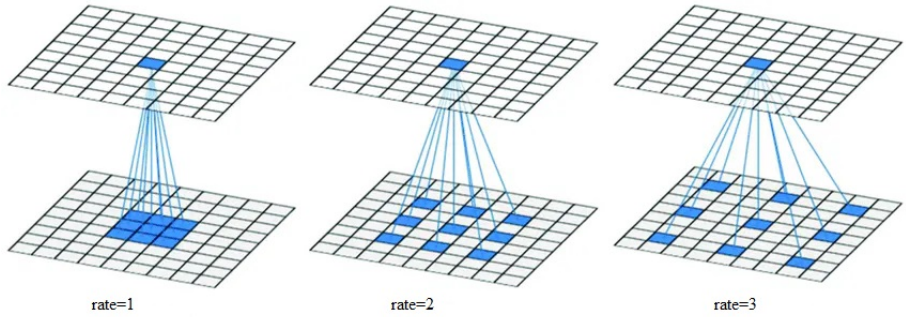
Atrous Spatial Pyramid

Mô-đun ASP[4] có thể khám phá đặc điểm của các vùng nhận thức khác nhau và tích hợp chúng thông qua phương pháp tích chập atrous [3] và phương pháp nối. Cấu trúc chi tiết được hiển thị trong hình 3.4. Tích chập atrous[3] trong ASP có thể giữ lại nhiều thông tin hình ảnh để cải

thiện độ chính xác của phân đoạn và mở rộng vùng nhận thức(receptive field) mà không sử dụng các lớp pooling, và nó có thể được định nghĩa bởi công thức sau:

$$y[i] = \sum_i^k = x[i + r * k] * w[k]$$

Trong đó, Với mỗi output y[i], tích chập atrous [3] sẽ tính toán bằng cách nhân tích chập bộ lọc w với feature map x. Với atrous rate r tương ứng với khoảng cách giãn cách giữa các dòng và cột bằng và được chèn bởi các giá trị 0, nếu atrous rate bằng r thì giãn cách các dòng và cột liên tiếp là r-1 dòng, cột. Minh họa cụ thể một atrous convolution với các tỉ lệ rate = 1,2,3 ở hình 3.3 Quá trình hoạt động của **Atrous Spatial Pyramid**

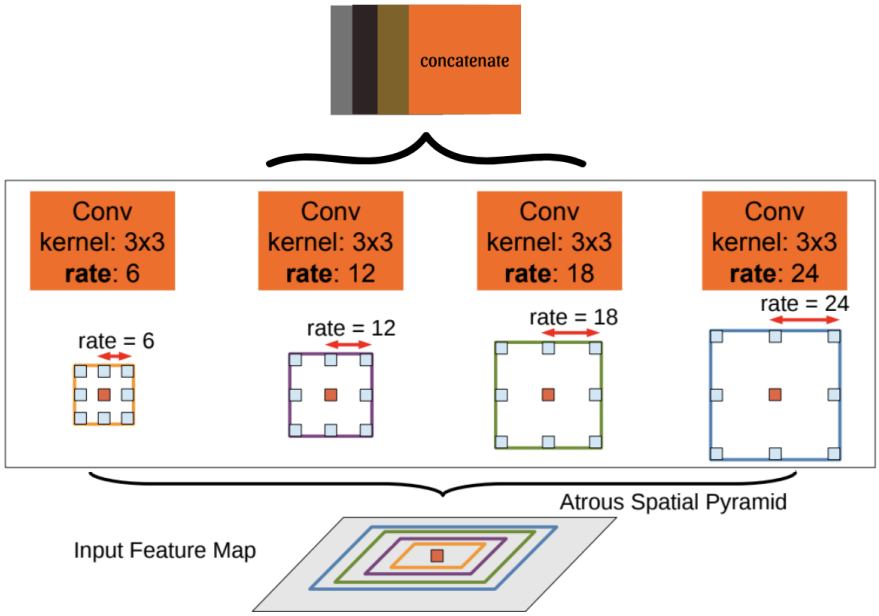


Hình 3.3: Kiến trúc Atrous [3]

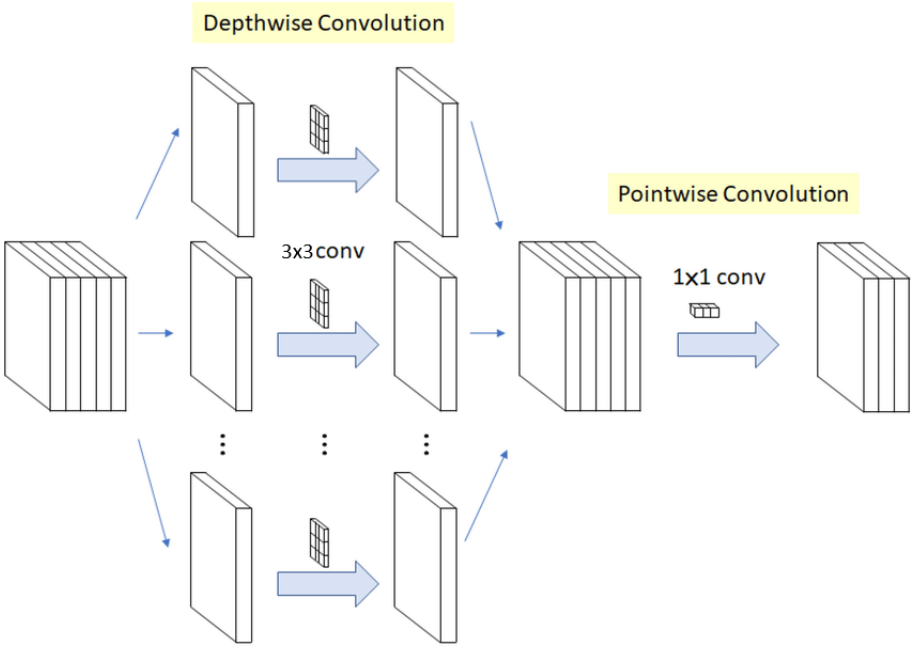
(ASP)[4] như sau: với đầu vào là một feature map, nó đi qua lần lượt các atrous convolution [3] với tỷ lệ rate = 6, 12, 18. Sự đa dạng về kích thước bộ lọc rate được xếp chồng lên nhau đã tạo thành một kim tự tháp bộ lọc atrous. Khi rate càng lớn, tầm nhìn của bộ lọc trên Input Feature Map càng lớn và giúp ta học được bối cảnh tổng quát (global context) tốt hơn. Với các bộ lọc kích thước nhỏ thì chúng ta sẽ học được bối cảnh cục bộ (local context) tốt hơn. Đồng thời, ASP[4] cũng giúp phát hiện vật thể ở nhiều kích thước khác nhau. Sau cùng, các đặc trưng tổng quát và cục bộ được trộn lẫn với nhau bằng cách nối các Feature Map output của các lớp atrous convolution[3] với nhau và trả về một score map chứa đầy đủ các thông tin của ảnh đầu vào.

Depthwise Separable Convolution [6]

Depthwise separable convolution (DSC)[6] là một biến thể của phép tích chập tiêu chuẩn, thường được sử dụng trong các mạng nơ-ron tích



Hình 3.4: Kiến trúc Atrous Spatial Pyramid[4]

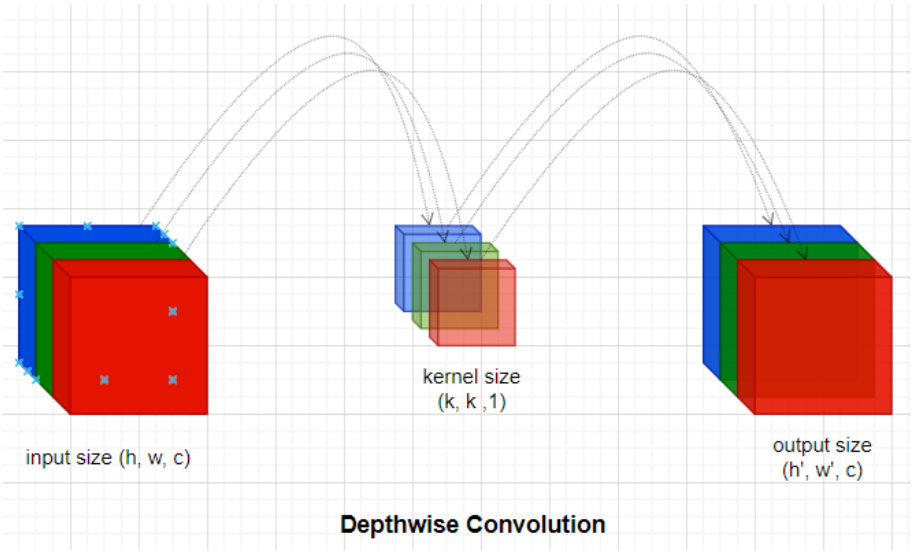


Hình 3.5: Các kênh đầu vào được tách riêng biệt, và mỗi kênh được thực hiện phép tích chập với bộ lọc không gian 3x3 (depthwise). Sau đó các kênh đã được tách được nối lại, và sau đó thực hiện phép tích chập 1x1 theo điểm (pointwise) để tạo ra số lượng kênh đầu ra mong muốn.[6]

chập. Nó được giới thiệu để giảm bớt độ phức tạp tính toán của các phép tích chập tiêu chuẩn, mà không đánh đổi quá nhiều hiệu suất của mô hình.

Trong một lớp convolution thông thường, chúng ta thực hiện tính toán trên cả không gian và kênh màu của ảnh trong một bước. Tuy nhiên, DSC tách bộ phận này thành hai bước riêng biệt là depthwise convolution và pointwise convolution để thực hiện 2 convolution, nhằm giúp giảm lượng phép tính cần phải thực hiện trong 1 convolution. Trước khi chúng ta đi chi tiết hơn về cách mà depthwise và pointwise convolution thực hiện, hãy cùng xem lại cách mà *convolution tiêu chuẩn (normal convolution)* hoạt động như thế nào và depthwise convolution hoạt động như nào ở hình 3.5.

Tích chập chiều sâu (Depthwise Convolution) Về cơ bản, tích chập sẽ chia khối input image có kích thước (h, w, c) thành những lát cắt ma trận theo độ sâu (chiều channels). Và thực hiện tích chập trên từng lát cắt như hình minh họa bên dưới: Trong **Depthwise Convolution**, mỗi



Hình 3.6: Cách hoạt động của tích chập theo chiều sâu

kênh đầu vào sẽ áp dụng một bộ lọc riêng biệt và không chia sẻ tham số với các bộ lọc khác. Điều này mang lại ba lợi ích chính cho mô hình:

- **Nhận diện đặc trưng**: Quá trình học và nhận diện đặc trưng được thực hiện riêng biệt cho mỗi bộ lọc. Điều này có nghĩa là, nếu các đặc trưng trên các kênh khác nhau, việc sử dụng các bộ lọc riêng biệt cho mỗi kênh sẽ giúp mô hình chuyên biệt hơn trong việc phát hiện các đặc trưng.

Ví dụ, nếu đầu vào là một hình ảnh màu RGB với ba kênh màu (Đỏ, Xanh lá cây, Xanh dương), thì mỗi kênh màu sẽ áp dụng một bộ lọc riêng biệt. Điều này giúp mô hình có thể học và nhận diện các đặc trưng riêng biệt trên mỗi kênh màu.

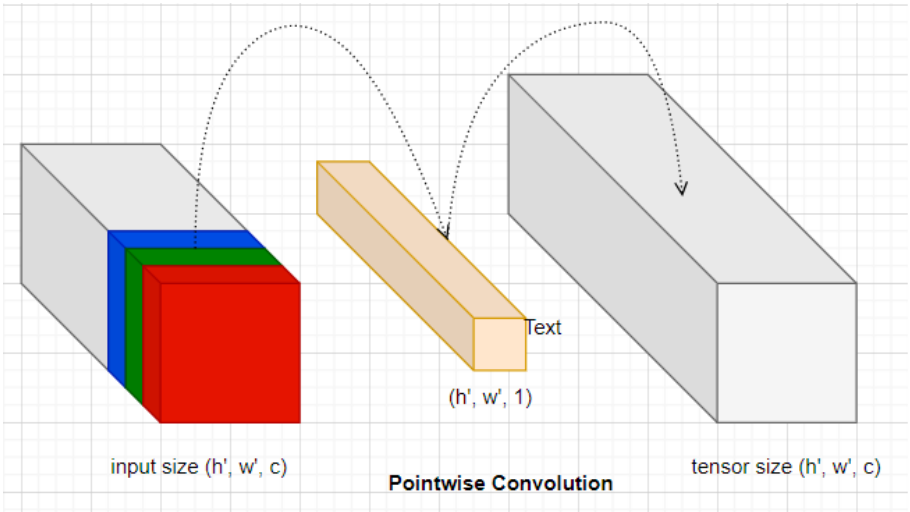
- **Giảm thiểu số lượng tham số**: Mỗi kênh đầu vào sẽ áp dụng một bộ lọc riêng biệt. Do đó, số lượng tham số cần sử dụng chỉ là $k \times k \times c$ (với k là kernel size và c là số chiều sâu channels). Trong khi một convolution thông thường sẽ có số lượng tham số cần sử dụng là $k \times k \times c'$ (với k là kernel size và c' là số chiều channels output cần thiết).

- **Tăng hiệu suất tính toán**: Mỗi kênh đầu vào sẽ áp dụng một bộ lọc riêng biệt. Do đó, để tạo ra một điểm pixel trên output, chúng ta chỉ cần thực hiện $k \times k$ phép tính. Trong khi đó, ở convolution thông thường, mỗi điểm pixel trên output được tạo ra từ $k \times k \times c$ phép tính, với c là số lượng kênh đầu vào. Điều này là do mỗi bộ lọc được áp dụng trên tất cả các kênh đầu vào. Vì vậy, giúp giảm đáng kể khối lượng tính toán, giúp tăng tốc độ tính toán và giảm bộ nhớ cần thiết cho mô hình.

Sau tích chập được concatenate lại theo độ sâu, output thu được là một tensor có kích thước $h' \times w' \times c$.

Tích chập điểm (Pointwise Convolution) Về cơ bản, tích chập điểm là một phép toán áp dụng một bộ lọc kích thước 1×1 lên từng điểm pixel của ảnh đầu vào. Điều này có tác dụng thay đổi độ sâu (số lượng kênh) của ảnh đầu vào mà không làm thay đổi kích thước chiều rộng và chiều cao. Cụ thể, nếu chúng ta muốn thay đổi độ sâu của ảnh đầu vào có độ sâu là c sang c'. Chúng ta sẽ áp dụng c' bộ lọc kích thước 1×1 và mỗi bộ lọc sẽ được áp dụng lên toàn bộ ảnh đầu vào và tạo ra một kênh đầu ra. Kết quả cuối cùng sẽ là một ảnh có kích thước giống ảnh đầu vào nhưng với độ sâu là c'. Như hình minh họa bên dưới:

Khi kết hợp cả hai phần trên, **Atrous Separable Spatial Pyramid (ASPP)**[4] (Hình 3.8) được tạo ra để mở rộng vùng nhận thức của mô hình bằng cách sử dụng các phép tích chập với các tỷ lệ **atrous rates**

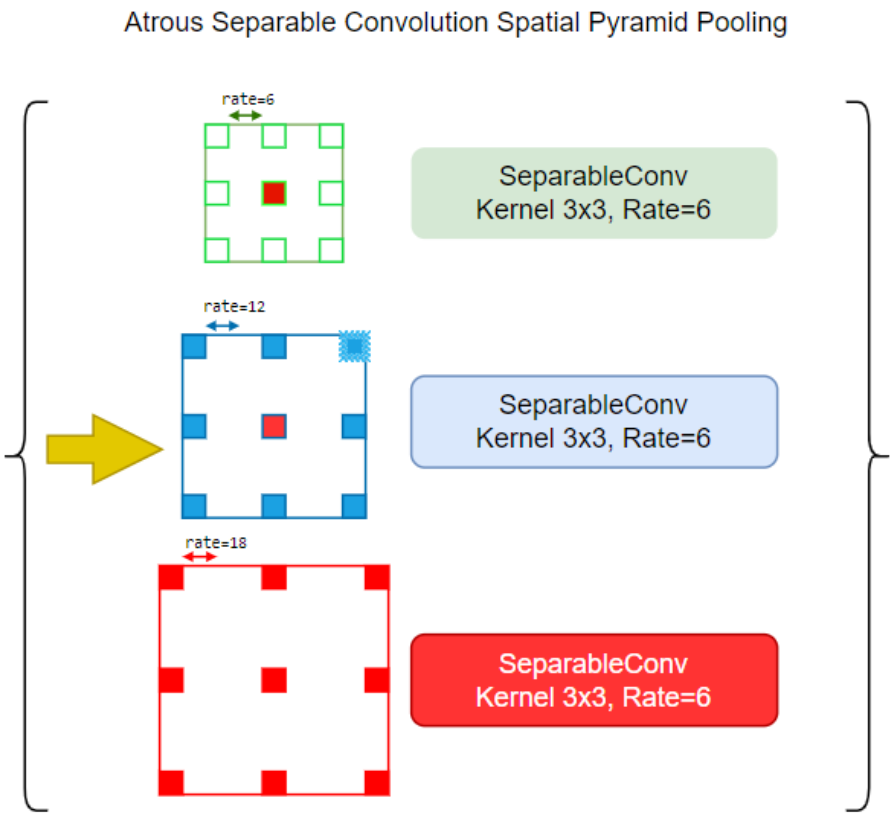


Hình 3.7: Cách hoạt động tích chập điểm

khác nhau, cho phép mô hình nhận thức được thông tin ở nhiều tỷ lệ khác nhau. Điều này cải thiện khả năng hiểu biết và dự đoán của mạng. ASPP[4] cũng kết hợp thông tin không gian và thông tin kênh bằng cách sử dụng các phép tích chập với kích thước kernel và số lượng kênh khác nhau. Điều này giúp mô hình học được các đặc trưng không gian và màu sắc ở nhiều tỷ lệ và độ phức tạp khác nhau, làm tăng độ chính xác và độ tin cậy của việc dự đoán. Ngoài ra, ASPP[4] giúp giảm thiểu mất mát thông tin trong quá trình giải mã. Các phép tích chập với **atrous rates** cho phép mô hình có khả năng hiểu biết thông tin chi tiết ở mức độ cao hơn, giúp giảm thiểu hiện tượng mất thông tin và tăng cường khả năng tái tạo chi tiết trong hình ảnh. Tóm lại, việc sử dụng ASPP[4] giúp mạng nơ-ron học sâu không chỉ mở rộng vùng nhận thức của mô hình, tăng cường thông tin không gian và khả năng tuỳ chỉnh mạnh mẽ mà còn giảm mất mát thông tin. Điều này giúp cải thiện hiệu suất và độ chính xác của mô hình trong các nhiệm vụ như phân đoạn hình ảnh.

3.1.3 Global Attention Upsample[22]

Trong lĩnh vực phân đoạn ngữ nghĩa hiện nay, có nhiều kiến trúc thiết kế khác nhau cho mô-đun giải mã. Ví dụ, **PSPNet** [41] hoặc **Deeplab** thường sử dụng phương pháp upsampling bilinear trực tiếp, đây có thể coi là một giải mã đơn giản. Trong khi đó, **DUC** [23] sử dụng phép tích chập kênh lớn kết hợp với việc điều chỉnh lại kích thước như một mô-đun giải



Hình 3.8: Kiến trúc Atrous Separable Spatial Pyramid Pooling(ASSPP)

mã một bước. Tuy nhiên, cả hai phương pháp trên đều thiếu thông tin từ các bản đồ đặc trưng cấp thấp ở các tỷ lệ khác nhau, và điều này có thể ảnh hưởng đến khả năng khôi phục vị trí không gian đến độ phân giải gốc.

Mô-đun **Global Attention Upsample (GAU)**[22] được phát triển để giải quyết những vấn đề của các phương pháp giải mã truyền thống như upsampling bilinear trực tiếp hoặc việc sử dụng các mạng encoder-decoder. GAU[22] sử dụng một kết hợp của phép tích chập và **Global Average Pooling**[24] để tạo ra một ngữ cảnh toàn cầu, từ đó giúp chọn lọc thông tin từ các đặc trưng cấp cao và hướng dẫn việc chọn chi tiết vị trí danh mục trong các đặc trưng cấp thấp.

Cụ thể, GAU thực hiện các bước sau:

1. Áp dụng phép tích chập 3x3 trên các đặc trưng cấp thấp để giảm kích thước kênh của bản đồ đặc trưng từ mạng neural convolutional networks (CNNs) [37].

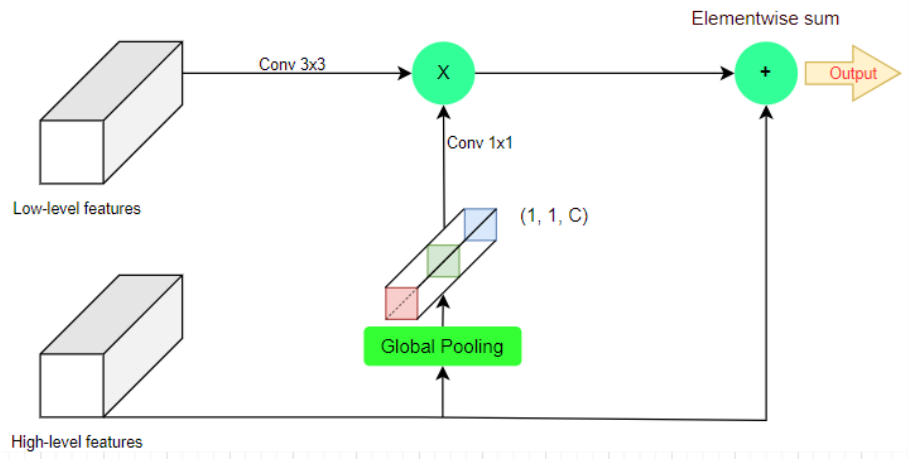
2. Tạo ra một ngữ cảnh toàn cầu từ các đặc trưng cấp cao thông qua phép tích chập 1x1 kết hợp với chuẩn hóa batch và hàm kích hoạt

phi tuyến tính ReLU.

3. Kết hợp ngữ cảnh toàn cầu với các đặc trưng cấp thấp bằng cách nhân chúng với nhau.

4. Dần dần tăng kích thước của các đặc trưng kết hợp này để tạo ra các bản đồ đặc trưng tỷ lệ khác nhau một cách hiệu quả.

Với các bước này, GAU[22] không chỉ cải thiện khả năng phục hồi vị trí không gian đến độ phân giải gốc, mà còn cải thiện hiệu suất của mô hình bằng cách sử dụng thông tin từ các đặc trưng cấp cao. Điều này giúp mô hình hiểu được ngữ cảnh toàn cầu và chọn lọc thông tin quan trọng để tăng cường khả năng phân loại và phân đoạn ảnh.



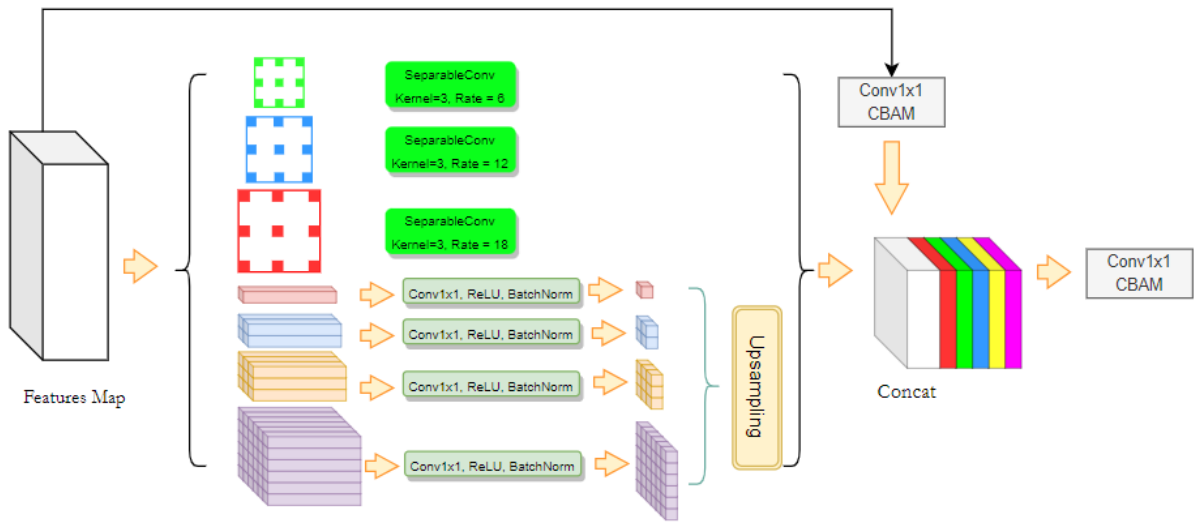
Hình 3.9: Kiến trúc Global Attention Upsample (GAU)[22]

3.2 Kiến trúc đề xuất

3.2.1 Mô-đun đề xuất

Mô-đun **Atrous Separable Spatial Pyramid Pooling (ASSPP)** được nhóm đề xuất là một cải tiến dựa trên mô-đun **Pyramid Pooling** của **PSPNet [41]**, với mục tiêu chính là cải thiện hiệu suất phân đoạn hình ảnh trên tập dữ liệu **Gleason 2019** [30] [19]. Minh họa cụ thể được trình bày trong hình3.10

Về cơ bản, mô-đun **ASSPP** bao gồm mô-đun **Pyramid Pooling** của PSPNet [41] và **Atrous Separable Pyramid**[4]. Đầu vào của mô-đun



Hình 3.10: Kiến trúc Atrous Separable Spatial Pyramid Pooling (AS-SPPM)

Pyramid Pooling là một bản đồ đặc trưng (features map), và qua mỗi lớp pyramid, features map này được xử lý với một tỷ lệ pooling khác nhau. Cụ thể, chúng ta sử dụng `AdaptiveAvgPool2d`[24] để thực hiện pooling theo tỷ lệ 1x1, 2x2, 3x3 và 6x6 trên không gian tính năng đầu vào. Đầu ra của các cấp độ khác nhau trong mô-đun pyramid pooling chứa bản đồ đặc trưng với các kích thước khác nhau. Để duy trì trọng lượng của đặc trưng toàn cầu, chúng ta sử dụng lớp convolution 1x1 sau mỗi cấp độ trong pyramid để giảm kích thước của biểu diễn ngữ cảnh xuống còn 1/N so với ban đầu với N là số tầng trong pyramid. Sau đó, chúng ta trực tiếp tăng kích thước của bản đồ đặc trưng có kích thước thấp để nhận được bản đồ đặc trưng cùng kích thước như bản đồ đặc trưng ban đầu thông qua nội suy bilinear [35].

Bên cạnh đó, mô-đun **Atrous Separable Pyramid** [4] được sử dụng để tăng cường khả năng nhận biết và phân loại đặc trưng của hình ảnh. Mô-đun này sử dụng pyramid các **atrous separable convolution** [6] với các tỷ lệ **Rate** (tỷ lệ atrous) khác nhau để tạo ra các bản đồ đặc trưng đa chiều, giúp mô hình có khả năng nhận biết các đặc trưng ở mức độ chi tiết và toàn cảnh. Cụ thể, mỗi lớp trong mô-đun Atrous Separable Pyramid [6] thực hiện một atrous separable convolution [6] với một tỷ lệ **Rate** cụ thể là 6, 12, 18. Atrous separable convolution [6] là sự kết hợp giữa **depthwise**

convolution và **pointwise convolution**, trong đó depthwise convolution được thực hiện với một kernel có tỷ lệ atrous, giúp mở rộng vùng nhận thức mà không làm mất thông tin không gian, và sau đó pointwise convolution được áp dụng để kết hợp thông tin từ các kênh khác nhau. Bằng cách sử dụng các atrous separable convolution [6] với các tỷ lệ **Rate** khác nhau, mô-đun Atrous Separable Pyramid [4] tạo ra một loạt các bản đồ đặc trưng có các receptive field có kích thước và phạm vi thông tin khác nhau. Điều này giúp mô hình có khả năng nhận biết các đặc trưng ở mức độ chi tiết (với các tỷ lệ atrous nhỏ) và cấu trúc toàn cảnh (với các tỷ lệ atrous lớn), tăng cường khả năng phân loại và phân đoạn hình ảnh. Việc thêm mô-đun **Atrous Separable Pyramid** [4] nhằm giúp giảm đi một số nhược điểm của mô-đun **pyramid pooling** thông thường. Mặc dù mô-đun pyramid pooling cho phép mô hình thu thập thông tin từ các vùng không gian với các tỷ lệ khác nhau, nhưng nó có thể dẫn đến mất mát thông tin do việc thực hiện pooling.

Trong khi đó, mô-đun Atrous Separable Pyramid [4] sử dụng **atrous separable convolution**[6], giúp mở rộng vùng nhận thức mà không làm mất thông tin không gian, giảm thiểu mất mát thông tin. Mô-đun Pyramid Pooling tập trung vào việc học các đặc trưng tổng quát từ ảnh, nhưng nó có thể không đủ mạnh để học được các đặc trưng tinh vi hoặc phức tạp. Tuy vậy, Atrous Separable Pyramid [4] cung cấp một phạm vi rộng hơn của các đặc trưng, từ các chi tiết nhỏ đến cấu trúc lớn của toàn bộ hình ảnh.

Hơn nữa, khi mô-đun Pyramid Pooling gặp khó khăn trong việc nhận diện các đối tượng nhỏ, Atrous separable convolution[6] trong mô-đun Atrous Separable Pyramid [4] cho phép mô hình nhận biết và phân loại các đặc trưng ở mức độ chi tiết và toàn cảnh của hình ảnh. Điều này giúp cải thiện độ chính xác của mô hình trong các nhiệm vụ như phân đoạn hình ảnh.

3.2.2 Kiến trúc đề xuất

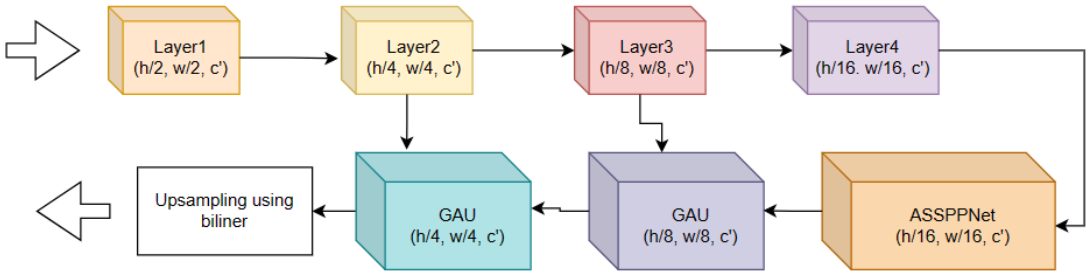
Với đề xuất **Atrous Separable Spatial Pyramid Pooling (AS-SPP)** và **Global Attention Upsample (GAU)**[22], như được minh họa

trong hình 3.11, nhóm đã sử dụng mạng **Resnext50-32x4d**[39] đã được huấn luyện trước trên tập **IMAGENET1K-V2** được cung cấp bởi thư viện **Torch**. Áp dụng chiến lược **atrous convolution** để trích xuất bản đồ đặc trưng. Cụ thể, thực hiện convolution với tỷ lệ **dilation=2** cho các khối **res4**, làm cho kích thước đầu ra của bản đồ đặc trưng từ **ResNet**[14] trở thành 1/16 của ảnh đầu vào, tương tự như **DeepLabv3+**[4] được gọi là dilation ResNet [14].

Mạng CNNs ResNet [14], được sử dụng làm kiến trúc gốc (backbone) để trích xuất đặc trưng từ hình ảnh đầu vào. Các khối dư thừa tích hợp sẵn trong ResNet [14] giúp tận dụng thông tin về loại hình của hình ảnh đầu vào từ một góc nhìn toàn cầu. Đặc biệt, các kết nối bỏ qua dài trong các khối dư thừa giúp thu thập thông tin về loại hình từ các lớp sâu hơn, cải thiện khả năng biểu diễn đặc trưng của mạng. Hơn nữa, như đã thảo luận trong [16], ResNet [14] là một mạng xuất sắc để trích xuất đặc trưng từ hình ảnh y học và vượt trội hơn so với phương pháp học máy truyền thống. Thay vì sử dụng **Resnet101** [14] như mô hình **PSPNet [41]** gốc, nhóm đã chọn mạng **CNNs ResNet50-32x4d**[39] để giảm số lượng tham số và năng lượng tính toán gần như một nửa, nhưng vẫn duy trì độ hiệu quả tương đương với **Resnet101**[14] trên cùng tập IMAGENET1K-V2.

Đặc trưng đầu ra của ResNet[14] được sử dụng làm đầu vào cho mô-đun Atrous Spatial Pyramid Pooling (ASPP). ASSPP là một mô-đun được thiết kế để thu thập thông tin toàn cầu từ các đặc trưng của ResNet [14]. Để tăng trường nhận biết của mạng PSPNet [41] bằng cách sử dụng phép tích chập phân tách theo chiều sâu (Depthwise Separable Convolution[6]) với ba tỷ lệ giãn khác nhau, đồng thời sử dụng bốn mô-đun pooling với các kích thước khác nhau được chia thành các vùng con khác nhau để tạo ra các biểu diễn đã được pooling cho các vị trí khác nhau Để duy trì trọng số của đặc trưng toàn cầu, sử dụng một lớp tích chập 1×1 và bản đồ đặc trưng được nâng lên cùng kích thước với các đặc trưng đầu vào bằng cách sử dụng nội suy song song. Sau mỗi kích thước pooling khác nhau để giảm chiều của biểu diễn ngữ cảnh xuống còn 1/N so với ban đầu, trong đó N là số lượng kích thước pooling khác nhau. Cuối cùng, các đặc trưng của các cấp độ khác nhau được nối lại để tạo ra tiền đề cảnh quan toàn cầu cuối cùng. Đáng chú ý, kích thước của tỷ lệ giãn phép tích chập phân tách

theo chiều sâu và kích thước của pooling toàn cầu có thể thay đổi. Nhóm đặt kích thước của tỷ lệ giãn dilation là 6, 12 và 18, qua đó giúp thu thập thông tin từ các vùng không gian với các tỷ lệ khác nhau. Sau khi thông qua các lớp này, các đặc trưng này được nâng mẫu lên kích thước bằng với đầu ra của ResNet [14]. Điều này giúp đảm bảo rằng thông tin không gian được giữ nguyên. Tiếp theo, các đặc trưng này được nối với các đặc trưng đầu ra gốc của ResNet [14] sau khi được áp dụng một cơ chế chú ý là Channel Attention Module (CBAM)[38]. CBAM[38] giúp mô hình tập trung vào các kênh quan trọng hơn trong quá trình học. Kết quả cuối cùng là một biểu diễn đặc trưng phong phú, vừa thu thập thông tin cục bộ từ đầu ra của ResNet [14], vừa thu thập thông tin toàn cầu từ việc nâng mẫu áp dụng trên các lớp pooling và tích chập. Điều này giúp mô hình có khả năng hiểu rõ hơn về cấu trúc không gian của dữ liệu đầu vào. Sau đó với bốn cấp độ pooling và 3 cấp độ tích chập cùng với features map của ResNet [14] áp dụng mô-đun chú ý, tất cả được nối lại với nhau và đi qua một lớp tích chập với kích thước kernel là 1 để giảm số chiều về kích thước 1/N so với đầu ra của ResNet [14]. Sau đó, chúng ta nâng mẫu lên gấp đôi hiện tại bằng cách sử dụng Global attention upsample lên đầu ra của mô-đun ASSPP. Lặp lại một lần nữa ta được kích thước features map bằng với 1/4 so với kích thước dài và rộng của ảnh đầu vào. Lúc này, nâng mẫu trực tiếp các bản đồ đặc trưng có chiều thấp để có cùng kích thước với bản đồ đặc trưng gốc thông qua phép nội suy tuyến tính. Cuối cùng, các đặc trưng từ các cấp độ khác nhau được nối lại để tạo thành *đặc trưng tổng hợp kim tự tháp cuối cùng*. Chi tiết kiến trúc mạng được minh họa chi tiết ở dưới đây:



Hình 3.11: Kiến trúc mạng đề xuất

3.2.3 Mạng lưới chi nhánh phụ trợ(Auxiliary Branch Network)[11]

Theo như [11], mạng phụ trợ này thực hiện các hoạt động như tích chập và lấy mẫu lên để xuất kết quả dự đoán. Hàm mất mát của mạng cốt lõi là L1 và hàm mất mát của mạng phụ trợ là L2, trong đó hàm mất mát L của toàn bộ mạng có thể được biểu diễn như sau:

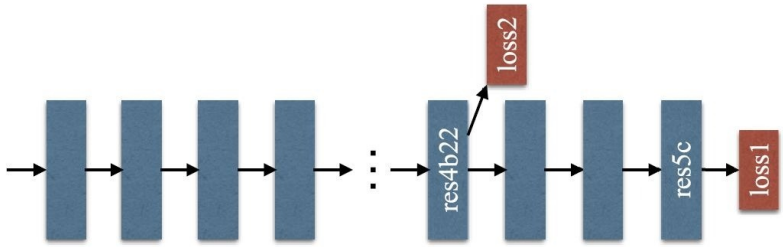
$$L = L_1 + \alpha * L_2$$

Trong mô hình mạng này, chúng tôi sử dụng hai hàm mất mát, L1 và L2, để tối ưu hóa việc học. Hệ số trọng số α được sử dụng để cân đối giữa hai hàm mất mát này, xét tham số $\alpha = 0.4$. Lý do để cài đặt tham số $\alpha = 0.4$ được tác giả của mô hình PSPNet [41] đưa ra dựa trên bảng so sánh .Cả hai hàm mất mát này đều là hàm mất mát cross-entropy.

Loss weight α	*mean Iou*	*Pixel Acc*
ResNet không cài α	35.82	77.07
ResNet50 với $\alpha = 0.3$	37.01	77.87
ResNet50 với $\alpha = 0.4$	37.23	78.0
ResNet50 với $\alpha = 0.6$	37.09	77.84
ResNet50 với $\alpha = 0.9$	36.99	77.87

Bảng 3.1: Bảng so sánh thông số cài đặt α trên ResNet50 [14]. Qua đó thấy được α có giá trị là 0.4 tối ưu nhất.

Điểm khác biệt so với hàm mất mát lan truyền ngược truyền thống là sử dụng hai cơ chế mất mát khác nhau, cho phép tối ưu hóa toàn bộ mạng, bao gồm tất cả các lớp tích chập. Điều này giúp tối ưu hóa việc học mạng mà không ảnh hưởng đến việc học của nhánh chính. Quá trình kết hợp thông tin cục bộ và toàn cầu để tránh mất thông tin, làm cho việc chẩn đoán và dự đoán bệnh trở nên đáng tin cậy hơn. Trong quá trình kiểm tra, thực hiện sử dụng nhánh chính, đã được tối ưu hóa tốt hơn, để dự đoán cuối cùng. Cụ thể minh họa ở hình sau 3.12 Cụ thể, có được bản đồ đặc trưng bằng cách đưa hình ảnh mỏng bệnh lý của ung thư tiền liệt tuyến vào ResNet [14]. Sau đó, bản đồ đặc trưng này được đưa vào ASSPP để học đặc trưng đa tỷ lệ. Cuối cùng, có thể nhận được kết quả dự đoán chính xác nhất.



Hình 3.12: Minh họa aux trong mạng resnet101. Với aux loss được thêm vào ngay block res4

3.2.4 Tổng kết

Trong phần trình bày trên, nhóm đã thực hiện nhiều cải tiến so với mô hình PSPNet [41] gốc. Mô hình đã sử dụng kiến trúc Resnext50-32x4d đã được huấn luyện trước trên tập IMAGENET1K-V2. Mô hình sử dụng mô-đun Atrous Spatial Pyramid Pooling (ASSPP) để thu thập thông tin toàn cầu từ các đặc trưng của ResNet [14]. Điều này giúp mô hình thu thập thông tin từ các vùng không gian với các tỷ lệ khác nhau. Đặc biệt, mô hình đã chọn mạng CNNs ResNet50-32x4d [39] để giảm số lượng tham số và năng lượng tính toán gần như một nửa, nhưng vẫn duy trì độ hiệu quả tương đương với Resnet101 trên cùng tập IMAGENET1K-V2. Với phương pháp cải tiến đã nêu, nhóm hy vọng có thể nâng cao hiệu suất và độ chính xác của mô hình cho bài toán phân đoạn tế bào yng thư từ hình ảnh bệnh lý.

Chương 4

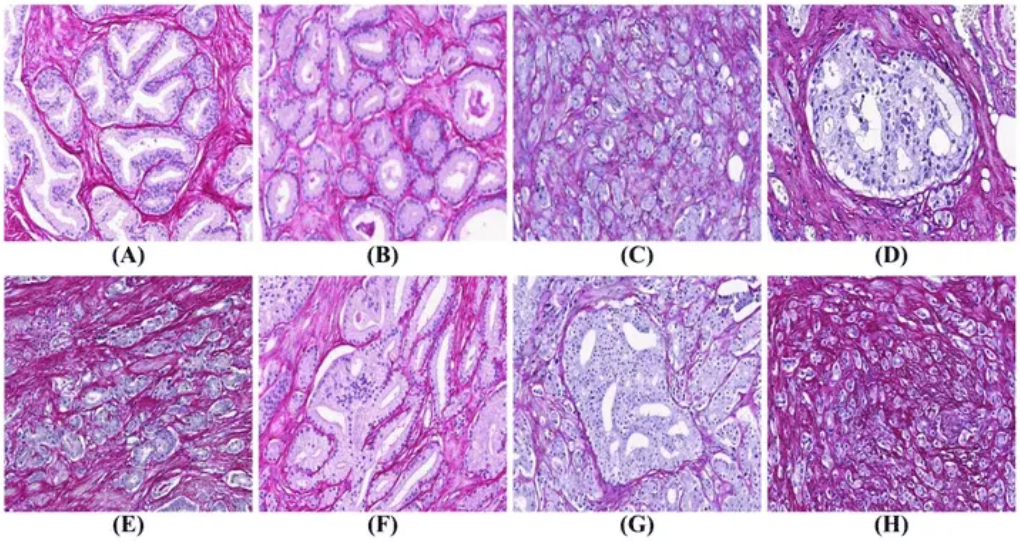
CÀI ĐẶT VÀ THỰC NGHIỆM

Từ các phương pháp cải tiến đã được đề xuất ở chương trước, chương này nhóm sẽ mô tả chi tiết quá trình và kết quả thực nghiệm khi cài đặt cải tiến đó vào mô hình PSPNet[41] cho bài toán phân đoạn ung thư tuyến tiền liệt dựa trên hình ảnh bệnh lý. Đầu tiên, nhóm sẽ giới thiệu các cơ sở dữ liệu, các độ đo được sử dụng trong quá trình thực nghiệm. Sau đó, môi trường thực nghiệm, các bước cài đặt thực nghiệm và quy trình thực nghiệm sẽ lần lượt được mô tả chi tiết. Cuối cùng, nhóm sẽ trình bày những kết quả thực nghiệm đã thu được, từ đó bàn luận để đưa ra nhận xét và đánh giá, đồng thời kết quả thực nghiệm của nhóm cũng sẽ được so sánh với các công trình nghiên cứu khác đã được công bố trong cùng bài toán và tập dữ liệu.

4.1 Chuẩn bị dữ liệu

Để thực nghiệm các phương pháp cải tiến đề xuất, nhóm tiến hành huấn luyện và đánh giá mô hình lần trên tập dữ liệu ảnh ung thư tuyến tiền liệt được cuộc thi Gleason MICCAI 2019[30] [19] đã cung cấp các chú thích từ sáu chuyên gia. Tuy nhiên, không phải tất cả các chú thích này đều hoàn chỉnh. May mắn thay, tổng hợp các chú thích từ tất cả các chuyên gia đã bao phủ đầy đủ dữ liệu. Điều này tạo ra thách thức trong việc sử dụng các chú thích này. Tất cả hình ảnh lõi TMA đều được chú thích là lành tính và theo các mẫu Gleason 3, 4 và 5, minh họa cụ thể trong hình 4.1. Sáu bác sĩ giải phẫu bệnh học đã vẽ các vùng đường viền kính (closed contours) trên hình ảnh giải phẫu bệnh và gán nhãn cho mỗi khu vực với các cấp độ. Mặc dù tất cả hình ảnh lõi TMA đều được chú thích chi tiết, nhưng không có hình ảnh nào hoàn chỉnh. Thậm chí, có hai hình ảnh trong chú thích của chuyên gia không tương ứng với bất kỳ hình ảnh

nào trong tập dữ liệu đào tạo. Tuy nhiên, để đào tạo mạng, ta cần một mẫu thực sự. Do đó, nhiệm vụ là hợp nhất các mẫu của sáu chuyên gia để đưa ra được mẫu chính xác nhất. Trong số sáu bác sĩ giải phẫu bệnh học, chỉ có bốn người đã gán nhãn cho tất cả các hình ảnh, trong khi hai người khác chỉ gán nhãn cho 191 và 92 hình ảnh. Để tận dụng tốt hơn tất cả các chú thích của chuyên gia, nhóm đã sử dụng thuật toán Major Voting để xây dựng lớp mask (ground truth label) cuối cùng thông qua việc hợp nhất tất cả các chú thích được cung cấp bởi sáu chuên gia.



Hình 4.1: Các cấp độ Gleason: (A) lành tính; (B) tuyến phôi hình thành tốt (Gleason cấp độ 3); (C) tuyến phôi hình thành kém (Gleason cấp độ 4); (D) cribriform (Gleason cấp độ 4); (E) tuyến phôi nhỏ kết hợp (Gleason cấp độ 4); (F) tuyến phôi lớn kết hợp (Gleason cấp độ 4); (G) ung thư trong ống dẫn (Gleason cấp độ 4); (H) tuyến phôi hình thành kém và tế bào đơn lẻ (Gleason cấp độ 4 và 5) [23]

Trong quá trình thực nghiệm, nhóm đã nhận ra rằng việc giữ nguyên kích thước ban đầu của hình ảnh với chiều dài và rộng là 5120*5120 (như hình 4.2) đã tạo ra một thách thức lớn. Khi thử nghiệm việc thu nhỏ kích thước xuống 512*512, nhóm đã phát hiện ra rằng điều này dẫn đến mất mát thông tin đáng kể. Cụ thể, trong một số hình ảnh có các cấp độ ung thư cao, các chi tiết quan trọng đã bị tiêu biến, làm giảm độ chính xác của dự đoán của mô hình. Để giải quyết vấn đề này, nhóm nhóm đã đưa ra một phương pháp mới: chia hình ảnh đầu vào thành các hình ảnh nhỏ

hơn. Cụ thể hơn, nhóm đã áp dụng phương pháp **sliding window** để cắt hình ảnh đầu vào thành các hình ảnh nhỏ hơn, mỗi hình ảnh có kích thước là 512*512 với độ chồng chất là 12 pixel. Qua quá trình này, nhóm đã tạo ra một tập dữ liệu ban đầu với tổng cộng 7165 hình ảnh cho việc huấn luyện mô hình. Điều này được thực hiện bằng cách di chuyển cửa sổ trượt qua hình ảnh đầu vào với bước trượt là 768 pixel. Tuy nhiên, không phải tất cả các hình ảnh đều hữu ích cho việc huấn luyện mô hình. Do đó, nhóm đã tiến hành loại bỏ các hình ảnh không chứa phần của tế bào hoặc chỉ chứa phần quả tế bào quá nhỏ và không rõ ràng. Bằng cách sử dụng ngưỡng mean và std để sàng lọc các ảnh không cần thiết. Sau quá trình này, tổng số lượng hình ảnh có thể sử dụng được cho quá trình huấn luyện giảm xuống còn 6457. nhóm đã chia tập dữ liệu này thành hai phần: tập huấn luyện và tập kiểm định. Tỉ lệ chia là 8:2, với 5156 hình ảnh được sử dụng cho bộ dữ liệu huấn luyện và 1301 hình ảnh cho bộ dữ liệu kiểm định.

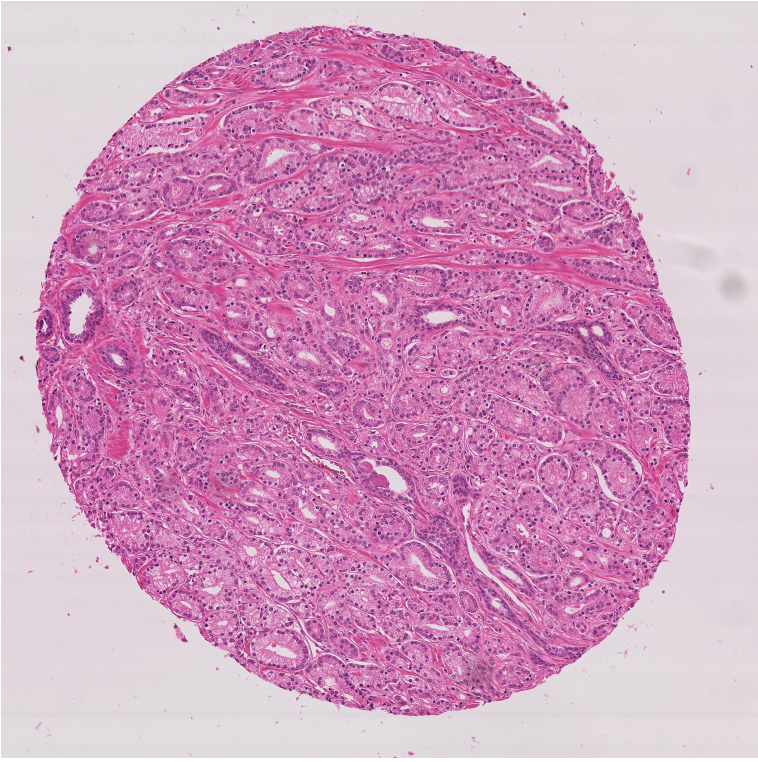
Hiện tại, vẫn chưa có sẵn bộ dữ liệu test đã được đánh nhãn cho cuộc thi. Do đó, nhóm chỉ có thể thực hiện việc dự đoán trên tập test, nhưng không thể tiến hành đánh giá chất lượng của các dự đoán này bằng cách so sánh với nhãn thực tế. Điều này dẫn đến việc nhóm không thể xác định được độ chính xác, độ nhạy, độ đặc trưng và các chỉ số đánh giá khác của mô hình trên tập test cho đến khi bộ dữ liệu test cụ thể được công bố.

4.2 Chiến lược đánh giá

Độ đo IoU (Intersection over Union)[17] hoặc còn được gọi là Jaccard Index, là một độ đo thường được sử dụng để đánh giá chất lượng của các mô hình phân đoạn hình ảnh. Nó tính toán tỷ lệ giữa diện tích phần chồng lấn của hai vùng dự đoán và thực tế chia cho tổng diện tích của chúng (minh họa như hình 4.3). Về mặt toán học mean IoU được tính như sau:

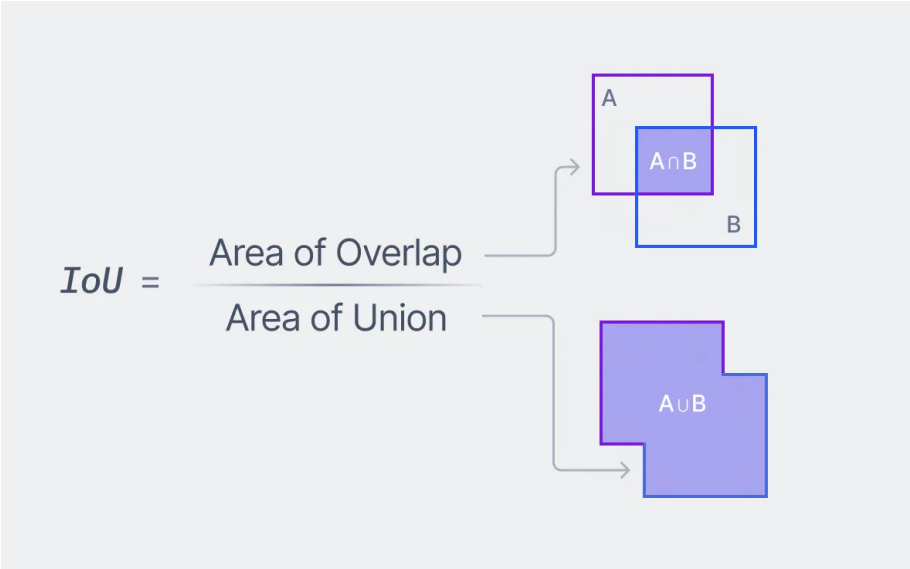
$$IntersectionoverUnion(IoU) = \frac{|A \bigcap B|}{|A| \bigcup |B|}$$

Mean IoU [17] là giá trị trung bình của IoU [17] trên tất cả các lớp hoặc các vùng phân đoạn trong tập dữ liệu. Nó được tính bằng cách lấy tổng IoU [17] của tất cả các phần chia cho số lượng lớp hoặc vùng phân



Hình 4.2: hình ảnh microarray mô (TMA) (tissue micro-array (TMA) image)

đoạn. Điểm IoU [17] sẽ cao nếu có nhiều sự chồng chéo giữa các hộp dự đoán và hộp thực tế. Ngược lại, sự chồng chéo ít sẽ dẫn đến một điểm IoU [17] thấp. Một điểm IoU [17] bằng 1 chỉ ra một sự khớp hoàn hảo giữa hộp dự đoán và hộp thực tế, trong khi một điểm số 0 đồng nghĩa với việc không có sự chồng chéo giữa các hộp. Mean IoU [17] cung cấp một cách tổng quan và tổng thể về hiệu suất của một mô hình phân đoạn bằng cách đánh giá khả năng phân loại chính xác của mỗi lớp hoặc vùng trong tập dữ liệu. Điều này giúp xác định mức độ chính xác của mô hình trên tất cả các loại đối tượng hoặc vùng quan trọng. Do đó, mean IoU [17] thường được chọn làm độ đo để đánh giá kết quả thực nghiệm trong hầu hết các nghiên cứu liên quan đến tác vụ phân đoạn ung thư. Vì vậy, nhóm quyết định sử dụng mean IoU [17] làm độ đo để đánh giá kết quả thực nghiệm của luận văn này.



Hình 4.3: Độ đo mean IoU [17]

4.3 Môi trường thực nghiệm

Khóa luận đã tiến hành huấn luyện và đánh giá các phương pháp cải tiến đề xuất trên máy tính cá nhân. Cấu hình chi tiết bao gồm : 1 CPU AMD Ryzen 7 3700X , có xung nhịp cơ bản là 3.6GHz và 4.4GHz khi ở trạng thái nâng cao. Với 8 nhân 16 luồng, 32MB ộ nhớ đệm. 1 GPU GeForce RTX 3070Ti với 8 GB ram GDDR6X và 6144 nhân CUDA được xây dựng trên kiến trúc NVIDIA Ampere. Ngoài ra, quá trình tiền xử lý và tăng cường dữ liệu được nhóm thực hiện trên laptop cá nhân với trình soạn thảo mã nguồn VS Code. Laptop được sử dụng mang cấu hình: CPU Core i7 10510U, 16GB ram và xung nhịp 2.3GHZ.

4.4 Cài đặt thực nghiệm

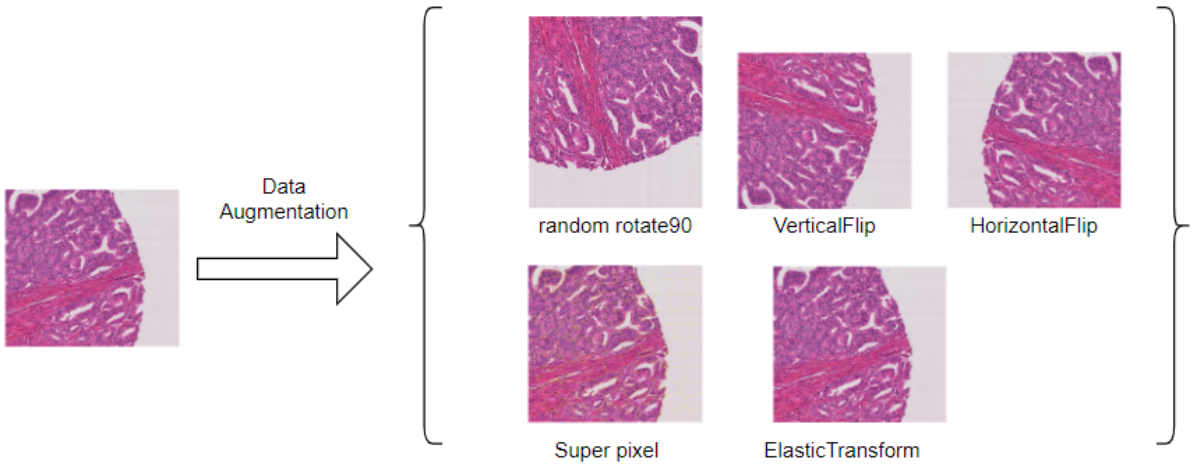
4.4.1 Ngôn ngữ và thư viện lập trình

Trong suốt quá trình thực hiện khoá luận, nhóm sử dụng Python làm ngôn ngữ lập trình chính. Trong đó, mô hình được xây dựng bằng thư viện pytorch. Bên cạnh đó, nhóm sử dụng thư viện Albumentations, đây là thư viện cung cấp một giao diện tăng cường hình ảnh ngắn gọn cho các tác vụ thị giác máy tính khác nhau, bao gồm phân loại đối tượng, phân đoạn và phát hiện đối tượng và ước lượng tư thế.

4.4.2 Tiền xử lý và tăng cường dữ liệu

Nhóm đã áp dụng một số kỹ thuật tăng cường dữ liệu phổ biến cho bộ dữ liệu huấn luyện, bao gồm: xoay ngẫu nhiên 90 độ, lật ngang, lật dọc và kỹ thuật tăng cường superpixel[13] và elastic transform[2]. Elastic Transform[2] là một kỹ thuật biến đổi hình ảnh mà trong đó hình ảnh được "co giãn" một cách ngẫu nhiên tạo ra sự đa dạng về dữ liệu. Superpixel[13] là một kỹ thuật trong xử lý hình ảnh, nơi mà một hình ảnh được chia thành các vùng nhỏ gọi là "superpixels". Mỗi superpixel bao gồm một nhóm các pixel liền kề có các đặc điểm màu sắc và cấu trúc tương tự nhau. Kỹ thuật này giúp giảm số lượng đơn vị cần xử lý trong hình ảnh, từ hàng triệu pixel xuống còn hàng nghìn superpixel, giúp tăng tốc độ xử lý và giảm nhu cầu về bộ nhớ. Đồng thời, việc sử dụng superpixel cũng giữ được thông tin không gian và cấu trúc của hình ảnh, điều này rất quan trọng trong nhiều tác vụ nhận dạng hình ảnh. Trong một nghiên cứu được công bố trên arXiv [40], phương pháp tăng cường dữ liệu dựa trên superpixel (SPDA) đã được đề xuất để huấn luyện các mô hình học sâu cho việc phân đoạn hình ảnh y sinh.

Ngoài ra, trước khi đưa vào huấn luyện, các ảnh trong thư mục dữ liệu cũng được trộn lẫn nhằm tăng thêm tính ngẫu nhiên cho trật tự của dữ liệu.



Hình 4.4: Các kỹ thuật tăng cường dữ liệu.

4.4.3 Các tham số và kỹ thuật thực nghiệm

Các tham số thực nghiệm bao gồm:

- Hệ số Learning rate ban đầu: 0.01.
- Kích thước batch: 16.
- Số lượng epoch: 200.
- Kích thước ảnh huấn luyện: 512x512.
- Hàm lỗi: Cross-Entropy Loss
- Thuật toán tối ưu hoá: SGD.

Các kỹ thuật tối ưu khi huấn luyện:

- Ký thuật learning rate policy: Điều chỉnh hệ số Learning rate theo từng epoch bằng cách nhân base learning rate với một tham số nhất định được tính theo công thức sau:

$$\text{current lr} = \text{base lr} * (1 - \frac{iter}{maxiter})^{power}), \text{ Với power} = 0.9$$

- kỹ thuật automatic mixed precision[1]: Đây là một kỹ thuật được sử dụng trong huấn luyện mô hình học máy để cải thiện hiệu suất và hiệu quả bộ nhớ. AMP hoạt động bằng cách sử dụng cả hai định dạng dữ liệu float16 (half precision) và float32 (single precision) trong quá trình huấn luyện, tận dụng lợi thế của cả hai: float16 để tăng tốc độ tính toán và giảm bộ nhớ, và float32 để duy trì độ chính xác số học. AMP tự động chọn định dạng dữ liệu phù hợp cho mỗi phép toán trong quá trình huấn luyện, giúp tối ưu hóa hiệu suất mà không làm giảm độ chính xác của mô hình. Nhóm đã sử dụng module torch.cuda.amp của thư viện Torch để áp dụng vào quá trình huấn luyện.

4.5 Quy trình thực nghiệm

Đầu tiên, nhóm đã tải các tập dữ liệu từ trang web của cuộc thi MICCAI 2019 về máy tính cá nhân. Sau đó, nhóm tiến hành tiền xử lý dữ liệu, bao gồm việc cắt ảnh thành các ảnh nhỏ và loại bỏ những ảnh không có nhiều ý nghĩa. Khi quá trình chuẩn bị dữ liệu hoàn tất, nhóm đã cài đặt thực nghiệm theo mô tả ở mục 4.4. Nhóm đã huấn luyện và đánh giá mô hình PSPNet[41] gốc trên từng tập dữ liệu để kiểm chứng khả năng của mô hình trong tác vụ phân đoạn hình ảnh tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Cuối cùng, nhóm đã cài đặt, huấn luyện và đánh giá các phương pháp cải tiến đề xuất trên từng tập dữ liệu để đánh giá hiệu quả của việc cải tiến.

4.6 Kết quả và nhận xét

4.6.1 Mô hình PSPNet[41]

Như đã đề cập ở mục 4.5, trước hết nhóm tiến hành huấn luyện và đánh giá mô hình PSPNet[41] gốc trên tập dữ liệu để kiểm chứng khả năng của mô hình trên tác vụ phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Kết quả thực nghiệm của mô hình PSPNet[41] và một số mô hình cơ sở khác sẽ được thể hiện ở Bảng

Mô hình	*mean IoU*	*Accuracy*
PSPNet[41]	0.661	0.88

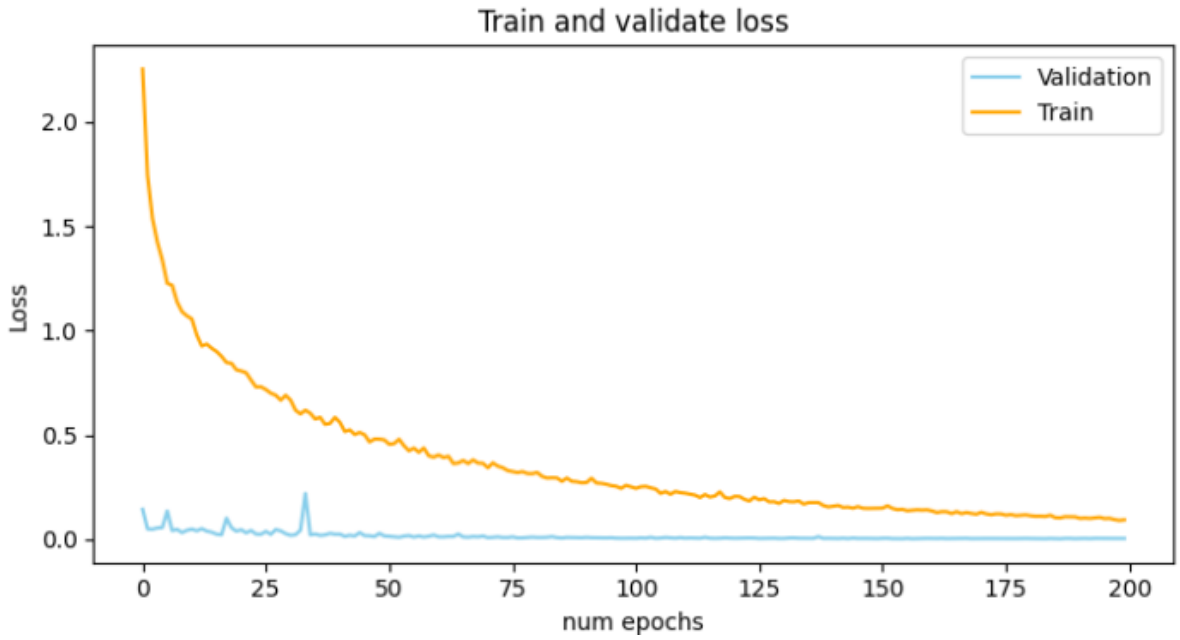
Bảng 4.1: Kết quả thực nghiệm của mô hình PSPNet[41] (độ đo mean IoU [17] và Accurancy)

Dựa trên kết quả thực nghiệm của mô hình PSPNet[41], nhóm đã nhận thấy rằng hiệu suất phân đoạn của mô hình này khi áp dụng cho bài toán phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý không đạt được như kỳ vọng. Mặc dù, PSPNet[41] là một mô hình tiên tiến và đã được chứng minh là có hiệu quả phân đoạn vượt trội so với các mô hình cơ sở khác. Tuy nhiên, khi áp dụng cho bài toán cụ thể này, kết quả không đạt

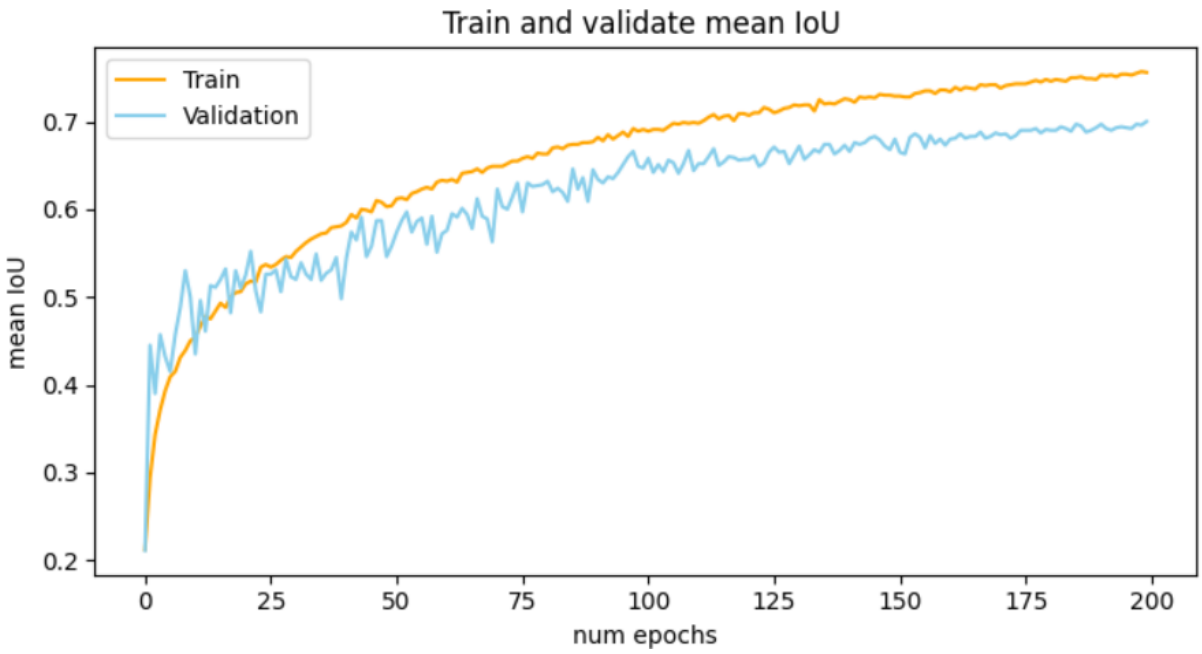
được như mong đợi. Điều này có thể là do kiến trúc ban đầu của mô hình chưa thực sự phù hợp với bài toán phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Mặc dù mô hình PSPNet[41] gốc không đạt được kết quả như mong đợi, nhóm vẫn tin rằng có thể cải thiện hiệu suất phân đoạn và khai thác tối đa tiềm năng của mô hình này. Điều này có thể được thực hiện thông qua việc áp dụng các phương pháp cải tiến đề xuất và điều chỉnh kiến trúc của mô hình. Bằng cách thực hiện những điều chỉnh, nhóm hy vọng sẽ có thể nâng cao hiệu suất của mô hình và đạt được kết quả tốt hơn trong việc phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Hy vọng rằng, thông qua quá trình này, nhóm sẽ có thể khai thác tối đa tiềm năng của mô hình và đạt được kết quả tốt nhất có thể.

4.6.2 Mô hình đề xuất

Như đã đề cập ở mục 4.5, để có thể chọn ra được kiến trúc tối ưu nhất của mô hình đề xuất, nhóm đã tiến hành huấn luyện và đánh giá lại mô hình sau mỗi lần tinh chỉnh lại kiến trúc của nó. Lịch sử huấn luyện của từng phiên bản khác nhau của mô hình đề xuất trên các tập dữ liệu được trực quan hoá ở Hình 4.5 và Hình 4.6



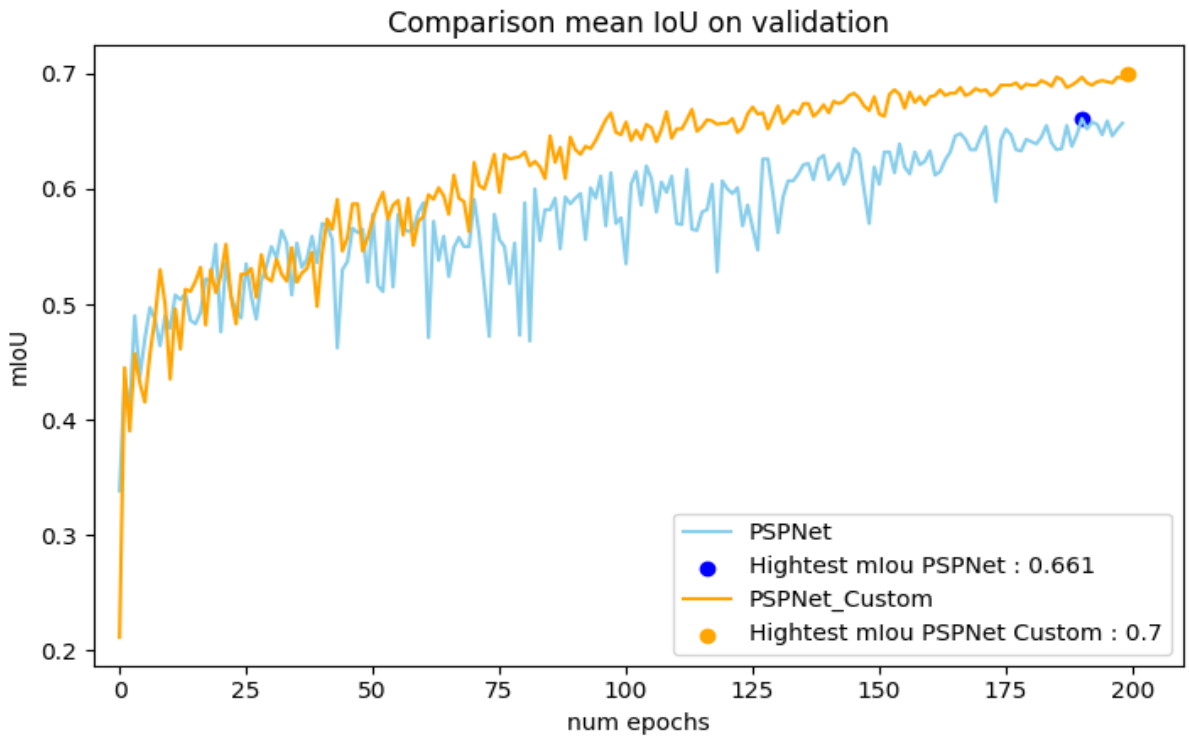
Hình 4.5: Biểu đồ thể hiện giá trị Loss trong quá trình huấn luyện mô hình



Hình 4.6: Biểu đồ thể hiện giá trị mean IoU [17] trong quá trình huấn luyện mô hình

Đầu tiên, giá trị mean IoU [17], một chỉ số đánh giá hiệu suất phân đoạn, trên cả tập huấn luyện và tập đánh giá có sự biến động qua từng epoch. Tuy nhiên, nhìn chung, giá trị này vẫn có xu hướng tăng dần, cho thấy quá trình huấn luyện đang diễn ra tốt. Mặc dù vậy, dấu hiệu của hiện tượng quá khớp (overfitting) bắt đầu xuất hiện ở các epoch sau, đây là điều cần lưu ý để điều chỉnh mô hình. Tiếp theo, giá trị mean IoU [17] trên tập huấn luyện tăng nhanh chóng trong những epoch đầu tiên, cho thấy mô hình đang học nhanh và dễ dàng tiếp thu các đặc trưng từ dữ liệu. Điều này là một dấu hiệu tích cực cho thấy mô hình có khả năng học tốt từ dữ liệu. Tuy nhiên, sau khoảng 125 epoch, tốc độ tăng của giá trị mean IoU [17] trên tập huấn luyện giảm đi khá nhiều. Điều này cho thấy quá trình huấn luyện đã hội tụ hoặc gần hội tụ. Điều này có thể chỉ ra rằng việc tiếp tục huấn luyện mô hình trong thời gian lâu hơn có thể không cải thiện đáng kể hiệu suất của mô hình. Nhóm cần xem xét việc điều chỉnh các tham số huấn luyện hoặc thay đổi cấu trúc mô hình để cải thiện hiệu suất.

Dựa trên kết quả thực nghiệm, nhóm nhận thấy mô hình đề xuất đều cải thiện đáng kể kết quả phân đoạn so với mô hình PSPNet[41] ban đầu.



Hình 4.7: So sánh kết quả thực nghiệm của mô hình đề xuất với mô hình gốc

Điều này cung cấp một bằng chứng mạnh mẽ để chứng minh rằng các thay đổi và cải tiến mà nhóm đề xuất cho mô hình PSPNet[41] là phù hợp và mang lại hiệu quả.

4.7 Đánh giá kết quả

Kết hợp các phương pháp cải tiến đã đề xuất, mô hình học sâu của nhóm đã đạt được kết quả tương đối cạnh tranh so với một số công trình nghiên cứu được công bố trước đây cho bài toán phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Cụ thể, kết quả phân đoạn của một số công trình nghiên cứu trên tập dữ liệu gleason 2019 challenge [30] [19].

Dựa trên bảng kết quả, nhóm có thể rút ra những nhận xét quan trọng về hiệu suất của mô hình đề xuất trong bài toán phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. ta thấy, mô hình PSPNet-improve đã cho thấy sự cải thiện đáng kể so với mô hình PSPNet[41] gốc. Về độ đo mean IoU, mô hình PSPNet-improve có mean IoU [17] là 0.7, cao hơn so với mô

Mô hình	*mean IoU*	*Accuracy*
PSPNet	0.661	0.88
PSPNet-imporove	0.7	0.913
MagNet(không thực hiện cắt ảnh) [15]	0.706	-

Bảng 4.2: Bảng so sánh kết của của mô hình trên tập dữ liệu Gleason [30] [19]

hình PSPNet[41] gốc (0.661). Điều này cho thấy mô hình PSPNet-improve có khả năng phân đoạn hình ảnh tốt hơn và cũng cho thấy độ chính xác cao hơn về accurancy (0.913) so với mô hình PSPNet[41] gốc (0.88). Điều này cho thấy mô hình PSPNet-improve có khả năng dự đoán chính xác hơn. Tuy nhiên, khi so sánh với mô hình MagNet[15] (không thực hiện cắt ảnh), mô hình PSPNet-improve có mean IoU [17] thấp hơn một chút (0.7 so với 0.706). Điều này có thể cho thấy rằng việc cắt ảnh có thể không luôn cần thiết để đạt được kết quả tốt, và có thể cần phải xem xét kỹ lưỡng hơn việc sử dụng kỹ thuật này tùy thuộc vào bộ dữ liệu và nhiệm vụ cụ thể. Nhìn chung, mô hình PSPNet-improve đã cho thấy sự cải thiện đáng kể so với mô hình PSPNet[41] gốc và có thể là một lựa chọn tốt cho việc phân đoạn hình ảnh trên tập dữ liệu Gleason [30] [19]. Tuy nhiên, do chưa có nhiều sự đầu tư cho phần cứng nên mô hình PSPNet-improve có thể vẫn chưa đạt được sự hiệu quả tối đa.

Chương 5

Kết luận và hướng phát triển

Trong phần này của khoá luận, nhóm sẽ tổng kết và đánh giá lại những gì nhóm đã học được và những thành tựu mà nhóm đã đạt được trong quá trình nghiên cứu và thực nghiệm. nhóm cũng sẽ trình bày một số phương hướng phát triển tiếp theo mà nhóm cho rằng có thể giúp cải thiện chất lượng công trình nghiên cứu này. Mục tiêu cuối cùng của nhóm là đưa công trình nghiên cứu này gần hơn với việc được áp dụng vào thực tế, nhằm tạo ra những giá trị thiết thực và có ý nghĩa cho cộng đồng nghiên cứu và xã hội nói chung

5.1 Kết luận

Trong khuôn khổ của khóa luận này, nhóm đã giới thiệu mô hình PSPNet-improve như một giải pháp mới và hiệu quả cho bài toán phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Thông qua các thử nghiệm thực tế, nhóm đã xác nhận được rằng mô hình PSPNet-improve không chỉ cải thiện đáng kể độ chính xác so với mô hình PSPNet gốc mà còn đạt được hiệu suất ấn tượng trong việc phân đoạn tế bào ung thư.

Khi kết hợp với các phương pháp cải tiến mà nhóm đề xuất, công trình nghiên cứu của nhóm đã tạo ra những kết quả cạnh tranh, thậm chí vượt trội so với một số nghiên cứu trước đây trong cùng lĩnh vực. Điều này không chỉ khẳng định giá trị và ý nghĩa của công trình nghiên cứu của nhóm đối với bài toán phân đoạn tế bào ung thư tuyến tiền liệt, mà còn góp phần quan trọng vào lĩnh vực phân đoạn ảnh y khoa nói chung.

5.2 Hướng phát triển

Trong quá trình thực hiện khóa luận, nhóm đã đề xuất và thực hiện một số phương pháp cải tiến trên mô hình PSPNet, tạo ra mô hình PSPNet-improve nhằm giải quyết bài toán phân đoạn tế bào ung thư tuyến tiền liệt từ ảnh bệnh lý. Mặc dù mô hình PSPNet-improve đã cho thấy kết quả khả quan, nhưng nhóm nhận thấy rằng vẫn còn một số hướng phát triển và cải tiến tiềm năng.

Đầu tiên, một trong những hướng cải tiến mà nhóm xem xét là việc sử dụng mô hình tiền huấn luyện. Mô hình tiền huấn luyện là một mô hình học sâu đã được huấn luyện trước trên một tập dữ liệu lớn và đa dạng, nhằm học được các đặc trưng cơ bản của dữ liệu. Việc sử dụng mô hình tiền huấn luyện có thể giúp tăng độ chính xác của mô hình, giảm thiểu thời gian và chi phí huấn luyện, và cải thiện khả năng chuyển đổi của mô hình cho các tác vụ khác nhau. nhóm nhận thấy rằng việc tiến hành tiền huấn luyện cho mô hình PSPNet-improve trên các tập dữ liệu lớn liên quan đến phân đoạn ảnh (như tập ImageNet) có thể là một hướng đi đáng xem xét để nâng cao hiệu quả phân đoạn và khả năng cạnh tranh của mô hình so với các công trình nghiên cứu khác.

Thứ hai, mô hình PSPNet-improve cần được cải thiện về tốc độ phân đoạn. Trong khóa luận này, nhóm đã tập trung vào việc cải tiến độ chính xác của kết quả phân đoạn, nhưng để mô hình có thể được ứng dụng vào thực tế, tốc độ phân đoạn cần phải đạt gần thời gian thực. Một số hướng đi có thể xem xét để cải thiện tốc độ phân đoạn bao gồm việc tối ưu hóa phần cứng, đơn giản hóa kiến trúc mô hình (ví dụ, giảm bớt số lượng tham số), và sử dụng các kỹ thuật tăng tốc huấn luyện như mixed-precision training hay distributed training.

Cuối cùng, mô hình PSPNet-improve có tiềm năng để được áp dụng vào các tác vụ phân đoạn ảnh y khoa khác, không chỉ riêng bài toán phân đoạn tế bào ung thư tuyến tiền liệt. Ví dụ, mô hình có thể được thử nghiệm trên các bài toán như phân đoạn khối u hay phân đoạn các cấu trúc khác trong cơ thể người. Việc thực hiện thêm các thử nghiệm như vậy sẽ giúp kiểm chứng tính đa dụng và khả năng tổng quát của mô hình, đồng thời mở rộng phạm vi ứng dụng của mô hình trong lĩnh vực y khoa.

Tài liệu tham khảo

Tiếng Anh

[1] *Automatic Mixed Precision for Deep Learning*. https://developer.nvidia.com/automatic-mixed-precision. Accessed: 2024-02-22. 2024.

[2] Castro, Eduardo, Cardoso, Jaime S., and Pereira, Jose Costa. "Elastic Deformations for Data Augmentation in Breast Cancer Mass Detection". In: *EMBC*. 2018. URL: https://web.fe.up.pt/~jsc/publications/conferences/2018EMecaBHI.pdf.

[3] Chen, Liang-Chieh et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs". In: *IEEE transactions on pattern analysis and machine intelligence* 40.4 (2017), pp. 834–848.

[4] Chen, Liang-Chieh et al. "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *ECCV*. 2018.

[5] Chen, Liang-Chieh et al. "Semantic image segmentation with deep convolutional nets and fully connected crfs". In: *arXiv preprint arXiv:1412.7062* (2014).

[6] Chollet, François. "Xception: Deep learning with depthwise separable convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[7] Cordts, Marius et al. "The Cityscapes dataset for semantic urban scene understanding". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[8] de Solorzano, Carlos Oscar. *Cell Tracking Challenge*. http://www.codesolorzano.com/celltrackingchallenge/Cell_Tracking_Challenge/Welcome.html. Accessed: 10-2-2024.

[9] Everingham, Mark et al. "The PASCAL Visual Object Classes VOC Challenge". In: *International Journal of Computer Vision (IJCV)* (2010).

[10] Everingham, Mark and Van Gool, Luc and Williams, Christopher K. I. and Winn, John and Zisserman, Andrew. *The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results*. `http://www.pascalnetwork.org/challenges/VOC/voc2011/workshop/index.html`. 2011.

[11] Fu, Siming et al. "AuxBranch: Binarization residual-aware network design via auxiliary branch search". In: *Pattern Recognition* 136 (2023), p. 109263. ISSN: 0031-3203. DOI: `https://doi.org/10.1016/j.patcog.2022.109263`. URL: `https://www.sciencedirect.com/science/article/pii/S0031320322007427`.

[12] Gupta, Saurabh et al. "Learning rich features from RGB-D images for object detection and segmentation". In: *European Conference on Computer Vision (ECCV)*. 2014.

[13] Hammoudi, Karim et al. "SuperpixelGridMasks Data Augmentation: Application to Precision Health and Other Real-world Data". In: *Journal of Healthcare Informatics Research* 6 (2023), pp. 442–460. URL: `https://hal.science/hal-03951542/document`.

[14] He, Kaiming et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[15] Huynh, Chuong et al. "Progressive Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[16] Iqbal, Saqib et al. "Prostate Cancer Detection Using Deep Learning and Traditional Techniques". In: *IEEE Access* 9 (2021), pp. 27085–27100. DOI: `10.1109/ACCESS.2021.3057654`.

[17] *Jaccard index*. `https://en.wikipedia.org/wiki/Jaccard_index`. Accessed: 2024-02-22. 2024.

[18] Jha, Debesh and al., et. "Kvasir-Instrument: Diagnostic and Therapeutic Tool Segmentation Dataset in Gastrointestinal Endoscopy". In: *MultiMedia Modeling*. Cham: Springer International Publishing, 2021, pp. 218–229.

[19] Karimi, D et al. "Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation". In: *IEEE journal of biomedical and health informatics* 24.5 (2019), pp. 1413–1426.

[20] Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012.

[21] Lam, L. and Suen, S. Y. "Application of majority voting to pattern recognition: an analysis of its behavior and performance". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 27.5 (1997), pp. 553–568. DOI: 10.1109/3468.618255.

[22] Li, Hanchao et al. "Pyramid attention network for semantic segmentation". In: *arXiv preprint arXiv:1805.10180* (2018).

[23] Li, Jie et al. "Real-time keypoints detection for autonomous recovery of the unmanned ground vehicle". In: *IET Image Processing* 14 (Feb. 2021). DOI: 10.1049/iet-ipr.2020.0864.

[24] Lin, Min, Chen, Qiang, and Yan, Shuicheng. *Network In Network*. 2014. arXiv: 1312.4400 [cs.NE].

[25] Lin, Tsung-Yi et al. "Microsoft COCO: Common Objects in Context". In: *European Conference on Computer Vision (ECCV)*. 2014.

[26] Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. "Fully convolutional networks for semantic segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[27] Maska, Martin et al. "A benchmark for comparison of cell tracking algorithms". In: *Bioinformatics* 30 (2014), pp. 1609–1617. DOI: 10.1093/bioinformatics/btu084.

[28] MIT Brain and Cognitive Sciences. *EM Segmentation Challenge*. http://brainiac2.mit.edu/isbi_challenge/. Accessed: 20-1-2024.

[29] Mottaghi, Roozbeh et al. "The role of context for object detection and semantic segmentation in the wild". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 891–898.

[30] Nir, G et al. "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts". In: *Medical image analysis* 50 (2018), pp. 167–180.

[31] Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer. 2015, pp. 234–241.

[32] Silberman, Nathan et al. "Indoor segmentation and support inference from RGBD images". In: *European Conference on Computer Vision (ECCV)*. 2012.

[33] Simonyan, Karen and Zisserman, Andrew. "Very deep convolutional networks for large-scale image recognition". In: *International Conference on Learning Representations (ICLR)*. 2015.

[34] Warfield, Simon K, Zou, Kelly H, and Wells III, William M. "Performance level estimation (STAPLE): an algorithm for the validation of image segmentation". In: *IEEE Transactions on Medical Imaging* 23.7 (2004), pp. 903–921. DOI: 10.1109/TMI.2004.828354.

[35] Wikipedia. *Nội suy — Wikipedia, Bách khoa toàn thư mở*. [Trực tuyến; truy cập 9-11-2023]. 2023. URL: https://vi.wikipedia.org/w/index.php?title=N%E1%BB%99i_suy&oldid=70872626.

[36] Wikipedia. *Prostate cancer — Wikipedia*. [Accessed: 9-11-2023]. 2023. URL: https://en.wikipedia.org/wiki/Prostate_cancer.

[37] Wikipedia contributors. *Convolutional neural network — Wikipedia, The Free Encyclopedia*. [Online; accessed 22-February-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Convolutional_neural_network&oldid=1209094961.

[38] Woo, Sanghyun et al. "Cbam: Convolutional block attention module". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[39] Xie, Saining et al. "Aggregated Residual Transformations for Deep Neural Networks". In: *CoRR* abs/1611.05431 (2016). arXiv: 1611.05431. URL: http://arxiv.org/abs/1611.05431.

[40] Zhang, Yizhe et al. *SPDA: Superpixel-based Data Augmentation for Biomedical Image Segmentation*. 2019. arXiv: 1903.00035 [cs.CV].

[41] Zhao, Hengshuang et al. "Pyramid scene parsing network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2881–2890.

[42] Zhou, Bolei et al. "Semantic understanding of scenes through the ADE20K dataset". In: *arXiv preprint arXiv:1608.05442* (2016).