

ĐẠI HỌC QUỐC GIA HÀ NỘI  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN  
KHOA TOÁN - CƠ - TIN HỌC



**HUS**  
VNU UNIVERSITY OF SCIENCE



---

**Báo cáo bài tập lớn**  
**Phân đoạn u não dựa trên dữ liệu ảnh MRI**

---

**Học phần: Một số vấn đề chọn lọc về Thị giác máy tính**

*Sinh viên thực hiện:*

Lê Mạnh Cương – 22001551

Nguyễn Tuấn Anh – 22001541

Phạm Quý Đô – 22001562

*Giảng viên:*

TS. Cao Văn Chung

*K67A3 – Khoa học Máy tính và Thông tin*

Ngày 7 tháng 1 năm 2026

## **Báo cáo bài tập lớn**

Phân đoạn u não dựa trên dữ liệu ảnh MRI

# Mục lục

<b>1</b>	<b>Giới thiệu bài toán</b>	<b>1</b>
<b>2</b>	<b>Bộ dữ liệu</b>	<b>4</b>
2.1	BraTS2020 . . . . .	4
2.2	Tiền xử lý dữ liệu . . . . .	5
2.2.1	Tiền xử lý dữ liệu 2D và tách lát cắt . . . . .	5
2.2.2	Tiền xử lý dữ liệu 3D . . . . .	6
2.3	Thông kê và phân tích bộ dữ liệu . . . . .	7
2.3.1	Phân tích phân bố thể tích của các vùng khối u (WT, TC, ET) . . . . .	7
2.3.2	So sánh đặc trưng tín hiệu giữa các modality MRI . . . . .	10
2.4	Phân chia tập dữ liệu . . . . .	11
<b>3</b>	<b>Phương pháp</b>	<b>13</b>
3.1	Quy trình tổng quan . . . . .	13
3.2	UNet . . . . .	13
3.3	UNet++ . . . . .	16
3.3.1	Skip connections được thiết kế lại trong UNet++ . . . . .	16
3.3.2	Deep Supervision trong UNet++ . . . . .	17
3.4	V-Net . . . . .	18
3.4.1	UNETR . . . . .	21
3.5	Swin UNet3D . . . . .	23
3.5.1	Tổng quan kiến trúc . . . . .	24
3.5.2	Biểu diễn đầu vào theo voxel patch và token . . . . .	25
3.5.3	Thiết kế theo stage: DownStage và Upstage . . . . .	25

3.6	Hàm mất mát . . . . .	26
3.6.1	Dice loss cho phân đoạn đa lớp . . . . .	26
3.6.2	Hàm mất mát kết hợp Dice + Cross-Entropy . . . . .	27
<b>4</b>	<b>Thực nghiệm</b>	<b>28</b>
4.1	Các độ đo đánh giá . . . . .	28
4.2	Chi tiết triển khai . . . . .	29
4.2.1	Mô hình UNet . . . . .	29
4.2.2	Mô hình UNet++ . . . . .	30
4.2.3	Mô hình VNet . . . . .	31
4.2.4	Mô hình UNETR . . . . .	36
4.2.5	Mô hình Swin UNet3D . . . . .	37
<b>5</b>	<b>Kết quả và thảo luận</b>	<b>38</b>
5.1	Kết quả . . . . .	38
5.2	So sánh các phương pháp . . . . .	47

# Danh sách bảng

5.1	Kết quả inference mô hình UNet trên tập test . . . . .	38
5.2	Kết quả inference UNet++ trên tập test . . . . .	39
5.3	Kết quả đánh giá các mô hình VNet trên các vùng WT, TC và ET. . . . .	41
5.4	Kết quả đánh giá các mô hình UNETR trên các vùng WT, TC và ET . . . . .	43
5.5	Kết quả inference mô hình Swin UNet3D trên tập test . . . . .	45
5.6	So sánh các phương pháp trên ROI WT . . . . .	47
5.7	So sánh các phương pháp trên ROI TC . . . . .	47
5.8	So sánh các phương pháp trên ROI ET . . . . .	47

# Danh sách hình vẽ

1.1	Các chuỗi MRI đa phương thức cùng với vùng phân đoạn bằng tay tương ứng của khối u não . . . . .	2
2.1	Các kết quả thống kê cho nhóm HGG . . . . .	8
2.2	Các kết quả thống kê cho nhóm LGG . . . . .	9
2.3	Minh họa một lát cắt trên 4 chuỗi MRI (Flair, T1, T1CE, T2) và nhãn phân đoạn của nó . . . . .	11
3.1	Sơ đồ tổng quan kiến trúc U-Net [?]	14
3.2	Kiến trúc tổng quan và skip connections của UNet++	17
3.3	Sơ đồ tổng quan kiến trúc V-Net	19
3.4	VNet downsample sử dụng tích chập 3D với kernel size $2 \times 2 \times 2$ và stride = 2	20
3.5	VNet upsampling sử dụng transposed convolution 3D với kernel size $2 \times 2 \times 2$ và stride = 2 . . . . .	21
3.6	Tổng quan về UNETR	22
3.7	Sơ đồ tổng quan kiến trúc UNETR	22
3.8	Kiến trúc tổng thể và các khối chính trong từng stage của Swin UNet3D	24
4.1	Quá trình huấn luyện mô hình UNet 2D: loss train và validation . . . . .	30
4.2	Biểu đồ thể hiện quá trình training với Unet++ . . . . .	30
4.3	Thử nghiệm 1 VNET: Quá trình training . . . . .	32
4.4	Huấn luyện VNet trên các patch 3D sử dụng hàm mất mát kết hợp DiceLoss + CELoss . . . . .	33
4.5	Huấn luyện VNet trên các patch 3D sử dụng hàm mất mát DiceLoss . . . . .	34
4.6	Huấn luyện VNet Multi-Head trên các patch 3D . . . . .	35
4.7	Huấn luyện VNet Multi-Encoder trên các patch 3D . . . . .	36

4.8	Biểu đồ quá trình training với UNETR . . . . .	37
5.1	Ví dụ kết quả phân đoạn của UNet 2D trên một lát cắt: ảnh MRI đầu vào, mask ground truth và mask dự đoán cho các vùng WT, TC, ET . . . . .	39
5.2	Kết quả dự đoán của mô hình trên lát cắt 88 của ca chụp 90 . . . . .	40
5.3	Kết quả phân đoạn của VNet trên lát cắt thứ 75 của ca chụp 11 . . . . .	42
5.4	Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 104 . . . . .	43
5.5	Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 193 . . . . .	43
5.6	Kết quả phân đoạn của UNETR trên lát cắt thứ 88 của ca chụp 091 . . . . .	44
5.7	Kết quả phân đoạn của UNETR trên lát cắt thứ 75 của ca chụp 011 . . . . .	44
5.8	Kết quả phân đoạn của Swin UNet3D trên lát cắt thứ 88 của ca chụp 091 . . . . .	45
5.9	Kết quả phân đoạn của Swin UNet3D trên lát cắt thứ 75 của ca chụp 011 . . . . .	46
5.10	Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 104 . . . . .	46
5.11	Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 193 . . . . .	46

# Chương 1

## Giới thiệu bài toán

U não là một dạng tăng sinh bất thường của các tế bào bên trong hộp sọ, gây ra sự ảnh hưởng trực tiếp đến hệ thần kinh trung ương. Khối u có thể là lành tính hoặc ác tính. Trong nhóm các khối u ác tính, glioma là dạng phổ biến nhất, xuất phát từ các tế bào thần kinh đệm và bao gồm 2 phân nhóm chính: Low-Grade Glioma (LGG) và và High-Grade Glioma (HGG) [...]. Mức độ nguy hiểm của 2 nhóm này khác nhau đáng kể: trong khi LGG tiến triển chậm với thời gian sống trung vị có thể đạt 11.6 - 11.7 năm, thì HGG (đặc biệt GBM) chỉ có thời gian sống trung vị khoảng 15 tháng [...]. Tỷ lệ tử vong cao cùng sự phát triển nhanh chóng của các khối u ác tính khiến việc phát hiện sớm, đánh giá kích thước, xác định ranh giới và lập kế hoạch điều trị trở thành nhiệm vụ then chốt trong thực hành lâm sàng.

Trong y khoa hiện đại, ảnh cộng hưởng từ MRI (Magnetic Resonance Imaging) là phương pháp hình ảnh lâm sàng được sử dụng rộng rãi nhất trong chẩn đoán u não do có ưu điểm: không bức xạ ion hóa, độ tương phản mô mềm cao và cung cấp được nhiều chuỗi xung khác nhau phản ánh những đặc tính sinh học riêng của mô não. Việc phân đoạn u não từ ảnh MRI đơn chuỗi là một nhiệm vụ khó khăn do cường độ ảnh có thể bị ảnh hưởng bởi partial volume effect hoặc các hiện tượng sai lệch trường (bias field artifacts) [?, ?]. Vì vậy, khối u não thường được chẩn đoán và đánh giá dựa trên các chuỗi MRI đa phương thức, bao gồm T1, T1 có tiêm chất tương phản T1CE, T2, và FLAIR. Khối u não bao gồm 3 phân vùng không chồng lấn:

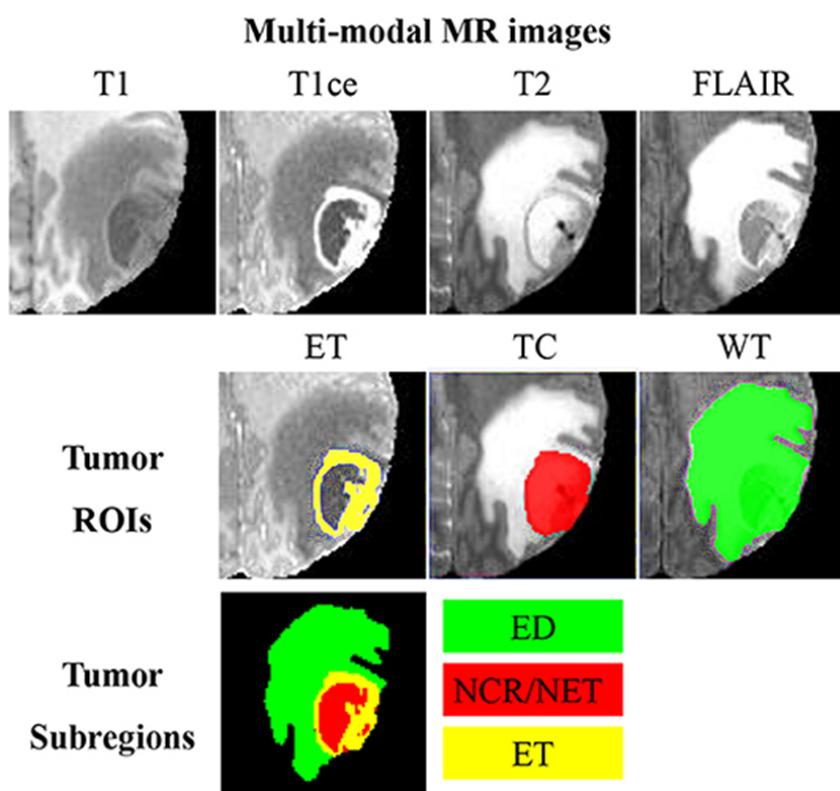
- Vùng u tăng sinh (Enhancing Tumor – ET): xuất hiện tăng tín hiệu mạnh trên T1CE.
- Vùng hoại tử hoặc mô không tăng quang (NCR/NET): thường giảm tín hiệu trên T1Gd so

với mô lành.

- Vùng phù quanh u (Edema – ED): thể hiện tăng tín hiệu đặc trưng trên FLAIR.

Các phân vùng này phản ảnh những đặc tính sinh học khác nhau của khối u. Từ 3 phân vùng trên, 3 vùng quan tâm (ROI) thường được sử dụng trong các tài liệu nghiên cứu có thể được tạo thành:

- Whole Tumor (WT) = NCR/NET + ED + ET  $\rightarrow$  toàn bộ vùng bất thường liên quan đến khối u.
- Tumor Core (TC) = NCR/NET + ET  $\rightarrow$  phần lõi khối u, không tính phù.
- Enhancing Tumor (ET)  $\rightarrow$  chỉ vùng tăng quang.



Hình 1.1: Các chuỗi MRI đa phương thức cùng với vùng phân đoạn bằng tay tương ứng của khối u não

Phân tích định lượng các ROI nói trên cung cấp những thông tin quan trọng phục vụ chẩn đoán bệnh, lập kế hoạch phẫu thuật và ước lượng tiên lượng, trong đó việc phân đoạn chính xác khối u và các ROI liên quan là vô cùng cần thiết. Các nhãn được vẽ thủ công bởi các chuyên gia chẩn đoán hình ảnh được xem là tiêu chuẩn vàng. Tuy nhiên, việc vẽ nhãn thủ công là quá tốn

công sức và mang tính chủ quan, khiến nó trở nên không khả thi trong hầu hết quy trình lâm sàng. Do đó, nhu cầu về các phương pháp phân đoạn u não tự động hoàn toàn là vô cùng cần thiết.

**Bài toán đặt ra:**

- **Đầu vào:** Ảnh MRI não dạng 2D hoặc 3D với đa chuỗi xung.
- **Đầu ra:** Bản đồ phân đoạn theo từng vùng u (ET, TC, WT).
- **Mục tiêu:** Xác định chính xác ranh giới u.

# Chương 2

## Bộ dữ liệu

### 2.1 BraTS2020

#### □ Mô tả dữ liệu:

Brain Tumor Segmentation Challenge 2020 (BraTS2020) là một trong những bộ dữ liệu chuẩn mực cho bài toán phân đoạn u não trên ảnh cộng hưởng từ (MRI) đa chuỗi. Bộ dữ liệu được xây dựng dựa trên MRI trước phẫu thuật của các bệnh nhân mắc u não dạng glioma, được thu thập từ nhiều trung tâm lâm sàng khác nhau nhằm đảm bảo tính đa dạng và khả năng tổng quát hóa của mô hình.

Bộ dữ liệu BraTS2020 bao gồm tổng cộng **369 trường hợp có nhän và 125 trường hợp không có nhän**. Đối với mỗi bệnh nhân, dữ liệu được cung cấp dưới dạng bốn chuỗi MRI khác nhau, bao gồm: **T1, T1ce, T2, Flair**. Ngoài bốn chuỗi ảnh đầu vào, mỗi trường hợp trong tập có nhän đi kèm một bản đồ phân đoạn (segmentation labelmap) chứa chú thích của chuyên gia lâm sàng. Các chú thích này phân biệt những vùng u khác nhau, bao gồm bốn giá trị nhän:

- 0 – background
- 1 – Vùng hoại tử hoặc nhän u không bắt thuộc tương phản (**NCR/NET**)
- 2 – Vùng phù quanh u (**ED**)
- 4 – Vùng u tăng tín hiệu sau tiêm thuốc tương phản (**ET**)

Tất cả dữ liệu ảnh MRI được cung cấp dưới định dạng NIfTI (.nii.gz) đã được tiền xử lý trước:

- Loại bỏ hộp sọ
- Chuẩn hóa không gian
- Resample về độ phân giải đẳng hướng  $1 \times 1 \times 1 \text{ mm}^3$
- Chuẩn hóa kích thước về dạng thể tích cố định  $240 \times 240 \times 155$  voxel.

## 2.2 Tiề̂n xử lý dữ liệu

Xây dựng hai nhánh tiền xử lý dữ liệu song song, tương ứng với 2 hướng tiếp cận mô hình:

- (i) Mô hình phân đoạn làm việc trên các lát cắt 2D
- (ii) Mô hình phân đoạn thể tích 3D.

Hai pipeline này được thiết kế sao cho nhất quán về mặt chuẩn hóa cường độ và mã hóa nhãn, đồng thời tối ưu chi phí tính toán cho từng cấu hình mô hình.

### 2.2.1 Tiề̂n xử lý dữ liệu 2D và tách lát cắt

Đối với bài toán phân đoạn 2D, dữ liệu thể tích MRI 3D được chuyển đổi thành tập các lát cắt 2D chuẩn hóa.

Quy trình gồm các bước chính sau:

#### ① Xác định vùng quan tâm toàn cục trên mặt phẳng 2D

Trước hết, sử dụng chuỗi T1 của tất cả các ca (cả có nhãn và không nhãn) để xác định vùng não trên mặt phẳng không gian hai chiều. Với mỗi thể tích, vùng não được suy ra từ các voxel có cường độ khác 0. Tập hợp các tọa độ này trên toàn bộ bệnh nhân được dùng để tìm một bounding box 2D tối thiểu bao phủ toàn bộ não trên mặt phẳng  $(x, y)$ .

Bounding box này sau đó được điều chỉnh lại thành hình vuông để khi resize ảnh, cấu trúc giải phẫu không bị biến dạng. Việc sử dụng một bounding box vuông, cố định cho mọi trường hợp đảm bảo rằng tất cả các lát 2D được cắt từ cùng một vùng giải phẫu tương đương, giảm diện tích nền không chứa não và tập trung mô hình vào vùng thông tin quan trọng.

## ② Chuẩn hóa cường độ bằng percentile

Đối với từng ca và từng chuỗi MRI, cường độ ảnh được chuẩn hóa độc lập. Chúng ta chỉ xét các voxel khác 0 (tương ứng với mô não), tính các ngưỡng percentile thấp và cao (ví dụ 1-percentile, 99-percentile), sau đó:

- Cắt ngưỡng cường độ về khoảng  $[p_{\min}, p_{\max}]$
- Normalize về khoảng  $[0, 1]$
- Các voxel nền ban đầu vẫn có giá trị 0.

Chuẩn hóa theo percentile giúp giảm ảnh hưởng của các giá trị ngoại lai, đồng thời đảm bảo phân bố cường độ ổn định hơn giữa các bệnh nhân và các chuỗi.

## ③ Cắt lát axial, crop và căn chỉnh hướng

Sau khi chuẩn hóa cường độ, mỗi thể tích 3D được cắt thành các lát cắt 2D theo mặt phẳng axial. Đối với mỗi lát, crop theo bounding box đã xác định ở bước 1, đảm bảo mỗi lát 2D chỉ chứa vùng não và loại bỏ phần lớn nền xung quanh.

Để đảm bảo hướng hiển thị nhất quán giữa các lát, các phép quay hoặc lật đơn giản có thể được áp dụng sau khi crop, sao cho não được “đứng thẳng” theo một quy ước chung. Bước này giúp thuận lợi hơn trong trực quan hóa kết quả và không làm thay đổi thông tin giải phẫu.

## ④ Resize ảnh và mã hóa nhãn

Các lát cắt sau khi được crop sẽ được resize về kích thước  $256 \times 256$  nhằm thống nhất kích thước ban đầu cho các mô hình 2D. Đối với ảnh cường độ (các chuỗi MRI), phép nội suy tuyến tính được sử dụng để bảo toàn cấu trúc cường độ. Đối với mặt nạ phân đoạn, nội suy lân cận gần nhất được sử dụng để tránh tạo ra các giá trị nhãn trung gian không hợp lệ.

Mặt nạ phân đoạn được mã hóa lại sao cho nhãn vùng u tăng tín hiệu ET (gốc là 4) được ánh xạ thành 3, tạo thành bộ nhãn  $\{0, 1, 2, 3\}$  liên tục.

### 2.2.2 Tiền xử lý dữ liệu 3D

Quy trình tiền xử lý dữ liệu 3D bao gồm các bước:

### **① Cắt giảm không gian quan tâm trong thể tích 3D**

Để loại bỏ vùng nền ít thông tin và tối ưu tài nguyên tính toán, chúng tôi áp dụng một cửa sổ cắt cố định trên hai trục không gian ( $x, y$ ), giữ nguyên toàn bộ chiều sâu. Vùng cắt được lựa chọn sao cho bao phủ đầy đủ toàn bộ não và khối u đối với mọi bệnh nhân, nhưng loại bỏ phần rìa ngoài chủ yếu là nền.

### **② Chuẩn hóa cường độ voxel trên mô não**

Tương tự tiền xử lý 2D, cường độ của các thể tích 3D được chuẩn hóa riêng cho từng trường hợp và từng chuỗi, nhưng theo kiểu thống kê toàn thể tích. Chỉ các voxel khác 0 (tương ứng với mô não) được dùng để tính thống kê.

Sử dụng chuẩn hóa z-score trên non-zero voxels, đưa cường độ về phân phối có trung bình xấp xỉ 0 và độ lệch chuẩn xấp xỉ 1. Các voxel nền được giữ nguyên bằng 0, giúp mô hình dễ dàng phân biệt giữa nền và mô não.

### **③ Chuẩn hóa nhãn phân đoạn trong thể tích 3D**

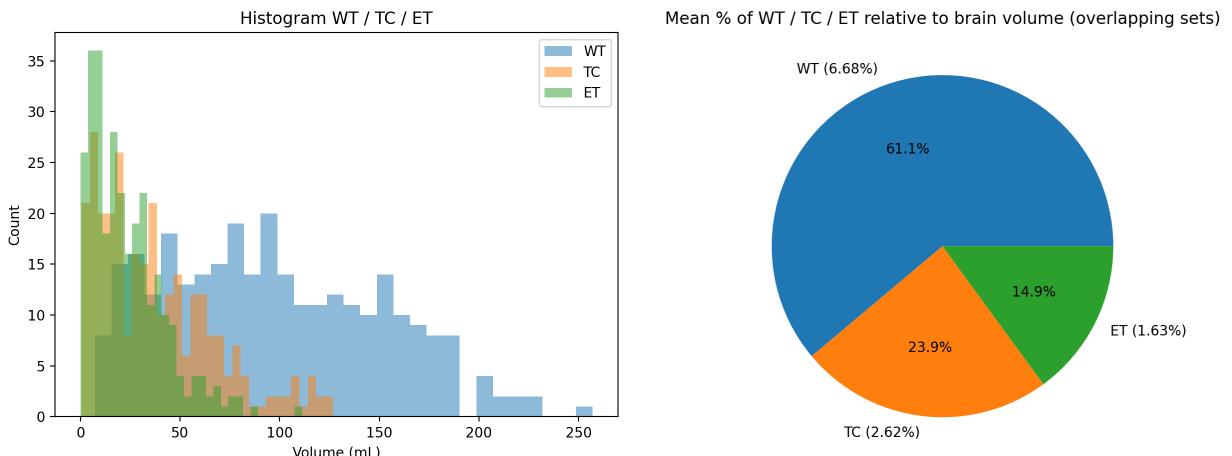
Mặt nạ phân đoạn 3D gốc sử dụng các nhãn 0,1,2,4. Tương tự như tiền xử lý 2D, thực hiện ánh xạ nhãn 4 (vùng u tăng tín hiệu) thành 3, thu được bộ nhãn 0,1,2,3.

## **2.3 Thông kê và phân tích bộ dữ liệu**

### **2.3.1 Phân tích phân bố thể tích của các vùng khối u (WT, TC, ET)**

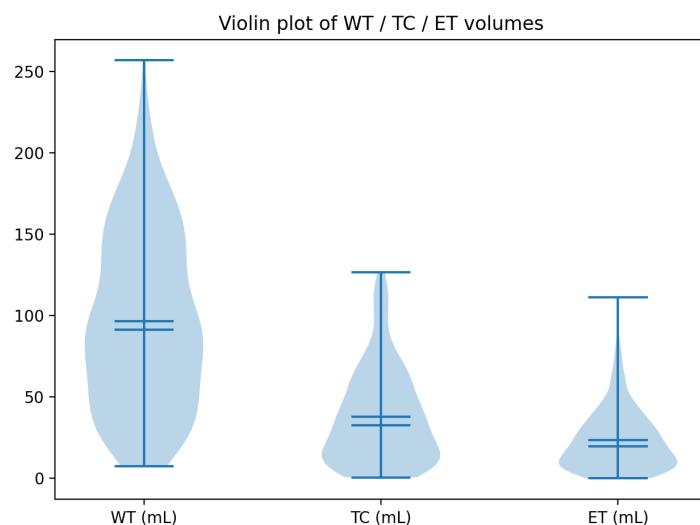
Để hiểu rõ hơn đặc tính hình thái của dữ liệu và đánh giá mức độ khác biệt giữa các nhóm u não, chúng tôi tiến hành phân tích thể tích của ba vùng quan trọng trong bộ dữ liệu BraTS2020, bao gồm Whole Tumor (WT), Tumor Core (TC) và Enhancing Tumor (ET). Phân tích được thực hiện riêng cho hai nhóm bệnh nhân High-Grade Glioma (HGG) và Low-Grade Glioma (LGG), nhằm làm rõ sự khác biệt về quy mô u giữa hai phân nhóm lâm sàng quan trọng này

#### **(a) Thông kê riêng cho nhóm HGG**



(a) Biểu đồ histogram thể tích các vùng WT / TC / ET cho nhóm HGG

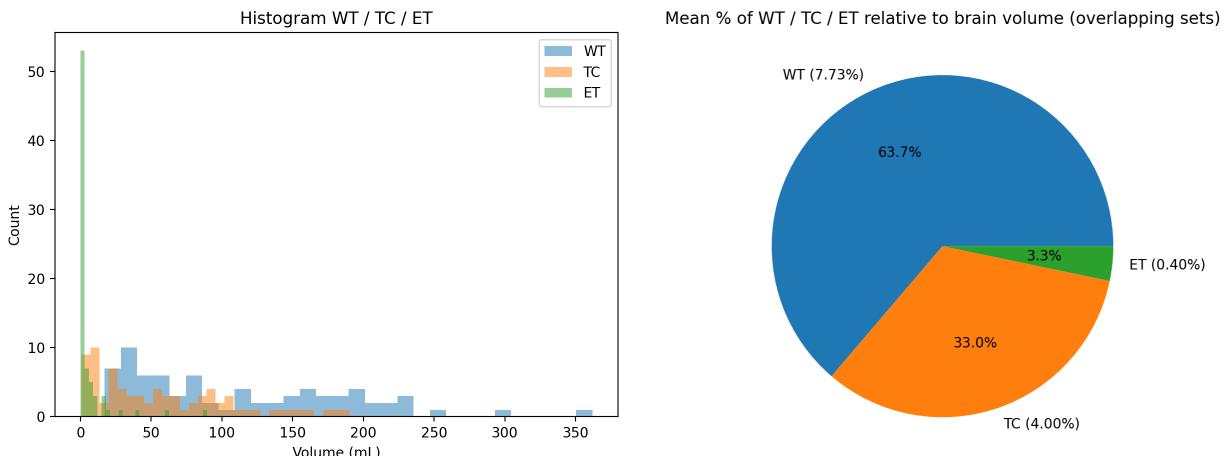
(b) Tỷ lệ phần trăm trung bình của WT / TC / ET so với thể tích não (nhóm HGG)



(c) Biểu đồ violin thể tích WT / TC / ET (nhóm HGG)

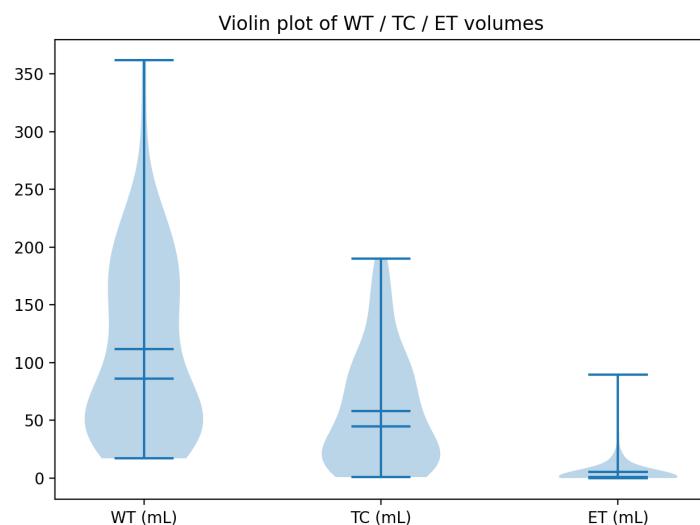
Hình 2.1: Các kết quả thống kê cho nhóm HGG

(b) **Thông kê riêng cho nhóm LGG**



(a) Biểu đồ histogram thể tích các vùng WT / TC / ET cho nhóm LGG

(b) Tỷ lệ phần trăm trung bình của WT / TC / ET so với thể tích não (nhóm LGG)



(c) Biểu đồ violin thể tích WT / TC / ET (nhóm LGG)

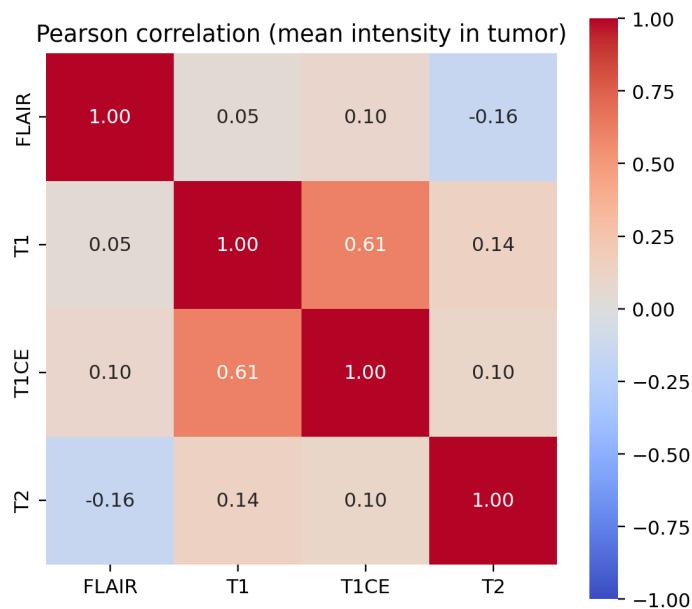
Hình 2.2: Các kết quả thống kê cho nhóm LGG

Kết quả phân tích cho thấy sự khác biệt rõ rệt về phân bố khối u giữa HGG và LGG, phù hợp với các nghiên cứu lâm sàng:

- HGG thường có kích thước WT, TC và ET lớn hơn, đồng thời phân bố thể tích trải rộng và không đồng nhất hơn.
- LGG thường có vùng ET rất nhỏ hoặc không xuất hiện, phản ánh mức độ tăng sinh thấp và tiến triển chậm.
- Sự khác biệt này có ý nghĩa quan trọng trong cả bài toán phân đoạn và bài toán phân loại mức độ ác tính, và cũng cho thấy cần thận trọng khi thiết kế mô hình để tránh thiên lệch về nhóm bệnh nhân có khối u lớn (thường là HGG).

### 2.3.2 So sánh đặc trưng tín hiệu giữa các modality MRI

Để đánh giá mức độ bối sung thông tin giữa các chuỗi MRI, tiến hành tính hệ số tương quan Pearson dựa trên cường độ trung bình của từng modality trong vùng khối u và trực quan hóa bằng heatmap.



Ma trận tương quan cho thấy các đặc điểm sau:

- T1 và T1CE có tương quan mạnh ( $r = 0.61$ )

Điều này phản ánh sự tương đồng về bản chất vật lý của hai chuỗi ảnh: T1CE (T1 có tiêm thuốc đồi quang từ gadolinium) thực chất là phiên bản tăng cường tín hiệu của T1, đặc biệt ở các vùng u tăng sinh mạch. Mức tương quan cao cho thấy hai chuỗi này có thông tin liên quan nhưng không hoàn toàn trùng lặp, do sự khác biệt giữa mô tăng sinh và mô không bắt thuốc.

- FLAIR có tương quan thấp với các modality còn lại

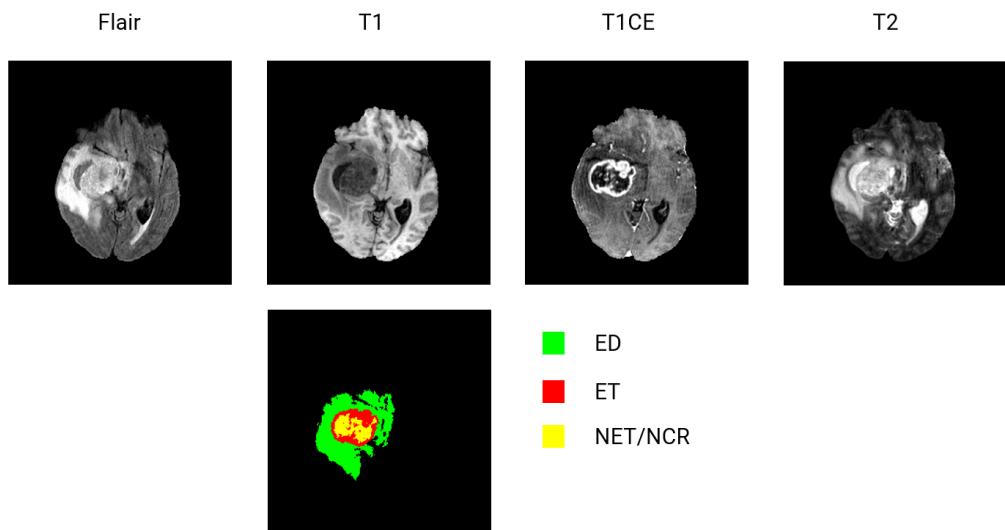
→ Phù hợp với đặc tính lâm sàng: Flair nhạy với vùng phù (ED), trong khi T1/T1CE phản ánh cấu trúc và mức độ tăng sinh mạch. Sự khác biệt này giúp FLAIR đóng vai trò bổ sung quan trọng trong việc mô tả ranh giới WT.

- T2 có tương quan rất thấp hoặc âm nhẹ với FLAIR ( $r = -0.16$ ).

Mặc dù FLAIR và T2 đều nhạy với mô chứa dịch, phản ứng của chúng trong vùng u có

thể khác nhau do sự ức chế tín hiệu dịch tự do trong FLAIR, làm giảm tương quan giữa hai chuỗi.

- T2 có tương quan yếu với T1 và T1CE (0.10–0.14)
  - T2 cung cấp thông tin ít trùng lặp với các chuỗi cấu trúc, góp phần tăng tính đa dạng thông tin đầu vào cho mô hình phân đoạn.



Hình 2.3: Minh họa một lát cắt trên 4 chuỗi MRI (Flair, T1, T1CE, T2) và nhãn phân đoạn của nó

Vì vậy, ta có thể kết luận:

- T1CE mang tính đại diện mạnh cho vùng ET, nhưng T1 vẫn cung cấp thông tin bổ sung.
- FLAIR và T2 bổ sung thông tin cho nhau, đặc biệt trong mô tả ranh giới WT và vùng phù.
- Sự đa dạng tín hiệu giữa các modality là cơ sở quan trọng cho việc sử dụng mô hình đa chuỗi (multi-modal learning), giúp mô hình học được các đặc trưng không gian–cường độ phức tạp của khối u glioma.
- Tương quan thấp giữa nhiều cặp modality cho thấy việc loại bỏ hoặc gộp các chuỗi ảnh không nên thực hiện đơn giản, bởi chúng đóng góp các đặc trưng khác biệt.

## 2.4 Phân chia tập dữ liệu

Sau khi hoàn tất bước tiền xử lý, dữ liệu được chia thành các tập con phục vụ huấn luyện, hiệu chỉnh và đánh giá mô hình. Việc phân chia được thực hiện trên mức bệnh nhân, thay vì trên mức

lát cắt, nhằm đảm bảo rằng dữ liệu từ cùng một bệnh nhân không bị rò rỉ giữa các tập train, validation và test.

Toàn bộ 369 ca có nhãn sẽ được chia thành 3 tập:

Tập	Tỷ lệ	Số ca chụp
Tập train	70%	258
Tập validation	15%	55
Tập test	15%	56

### **Phân tầng và đảm bảo tính công bằng khi chia dữ liệu**

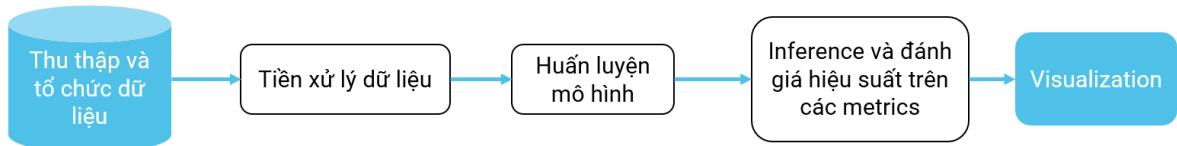
Để đảm bảo các tập train, validation và test có phân bố đặc trưng khối u tương đồng, quá trình chia dữ liệu được thực hiện theo chiến lược phân tầng thay vì chia ngẫu nhiên. Mỗi trường hợp bệnh nhân được mô tả bằng một số đặc trưng tổng hợp rút ra từ mặt nạ phân đoạn 2D, bao gồm: có/không có khối u, có/không có vùng u tăng tín hiệu (Enhancing Tumor), tổng diện tích khối u và số lát chứa u. Đồng thời, nhãn độ ác tính (HGG, LGG hoặc Unknown) được gắn cho từng ca dựa trên metadata.

Từ các đặc trưng này, chúng tôi xây dựng nhóm phân tầng kết hợp ba yếu tố: độ ác tính, sự hiện diện của vùng Enhancing Tumor và nhóm kích thước khối u (dựa theo quantile). Trong trường hợp một số nhóm có quá ít mẫu, các nhóm này được tự động gộp theo thứ tự giảm dần độ chi tiết để duy trì sự cân bằng. Việc chia train/validation/test được thực hiện ở mức bệnh nhân, với hạt giống ngẫu nhiên cố định nhằm đảm bảo tính tái lập và tránh rò rỉ dữ liệu. Cách tiếp cận này giúp duy trì phân bố tương đối đồng đều giữa các tập và đảm bảo kết quả đánh giá phản ánh đúng khả năng tổng quát hóa của mô hình.

# Chương 3

## Phương pháp

### 3.1 Quy trình tổng quan



### 3.2 UNet

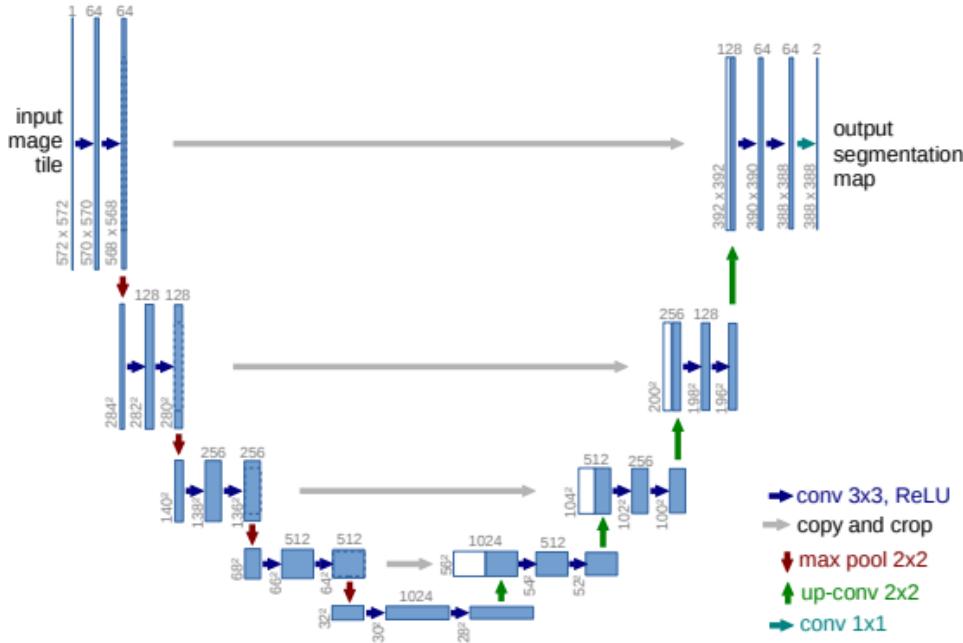
U-Net là một kiến trúc mạng nơ-ron tích chập (CNN) được đề xuất năm 2015 cho bài toán phân đoạn ảnh. Khác với các mô hình phân loại ảnh chỉ gán một nhãn cho toàn bộ bức ảnh, U-Net được thiết kế để gán nhãn cho từng điểm ảnh (pixel-wise), tức là ánh xạ ảnh đầu vào kích thước  $H \times W$  (với  $C$  kênh) sang một bản đồ nhãn cùng kích thước không gian, trong đó mỗi pixel được gán vào một trong  $K$  lớp (bao gồm cả nền và các vùng khối u như WT, TC, ET).

Điểm đặc trưng của U-Net là kiến trúc đối xứng dạng chữ U gồm hai nhánh:

- **Nhánh encoder (contracting path)** ở bên trái: trích xuất đặc trưng và thu gọn dần kích thước không gian.
- **Nhánh decoder (expansive path)** ở bên phải: khôi phục lại độ phân giải không gian và sinh bản đồ phân đoạn đầu ra.

Hai nhánh được kết nối với nhau bằng các *skip connection*, giúp kết hợp thông tin chi tiết mức

thấp (biên, texture) với thông tin ngữ nghĩa mức cao (ngữ cảnh toàn cục). Sơ đồ tổng quan kiến trúc U-Net được minh họa trong Hình 3.1.



Hình 3.1: Sơ đồ tổng quan kiến trúc U-Net [?]

**Nhánh encoder (contracting path).** Nhánh encoder của U-Net có cấu trúc tương tự một CNN dùng cho phân loại: mỗi tầng (level) gồm hai lớp tích chập  $3 \times 3$  liên tiếp (thường kèm BatchNorm và hàm kích hoạt ReLU), sau đó là một lớp gộp cực đại (MaxPooling) kích thước  $2 \times 2$  với stride 2 để giảm một nửa kích thước không gian:

$$(H, W, C) \xrightarrow{\text{Conv-BN-ReLU} \times 2} (H, W, C') \xrightarrow{\text{MaxPool } 2 \times 2} \left( \frac{H}{2}, \frac{W}{2}, C'' \right).$$

Sau mỗi lần down-sampling, số kênh đặc trưng thường được tăng gấp đôi (ví dụ  $32 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512$ ), giúp mạng học được các đặc trưng ngày càng trừu tượng hơn, đồng thời mở rộng *receptive field* để thu được nhiều ngữ cảnh toàn cục.

**Lớp bottleneck.** Tại đáy của chữ U là khối *bottleneck*, cũng gồm hai lớp tích chập  $3 \times 3$  liên tiếp. Đây là nơi tập trung các đặc trưng có mức trừu tượng cao nhất, đại diện cho nội dung toàn cục của lát cắt 2D. Trong mô hình UNet 2D sử dụng trong báo cáo này, bottleneck là tầng có số kênh lớn nhất và đóng vai trò cầu nối giữa encoder và decoder.

**Nhánh decoder (expansive path) và skip connection.** Nhánh decoder có nhiệm vụ khôi phục dần độ phân giải không gian để tạo bản đồ phân đoạn đầu ra. Mỗi level trong decoder

thường gồm:

1. Một bước *upsampling* (ví dụ dùng transposed convolution với kernel  $2 \times 2$  và stride 2, hoặc nội suy bilinear kết hợp tích chập) để phóng to kích thước không gian lên gấp đôi.
2. Phép **concatenate** với feature map tương ứng ở cùng mức trong encoder thông qua skip connection.
3. Hai lớp tích chập  $3 \times 3$  (có thể kèm BatchNorm + ReLU) để trộn và tinh chỉnh đặc trưng sau khi ghép.

Các skip connection giúp mô hình tận dụng lại đặc trưng biên và chi tiết hình học ở độ phân giải cao từ encoder (thường bị mất một phần khi pooling) và kết hợp chúng với thông tin ngữ nghĩa ở decoder. Nhờ vậy, U-Net vừa nắm bắt được bối cảnh toàn cục, vừa giữ được ranh giới khối u một cách sắc nét, đặc biệt hữu ích trong phân đoạn các vùng nhỏ như TC và ET.

**Lớp đầu ra và bản đồ nhãm.** Sau khi đi qua toàn bộ các tầng của decoder, feature map cuối cùng có kích thước không gian  $H \times W$  tương ứng với kích thước lát cắt đầu vào, nhưng với số kênh đặc trưng  $C_{\text{feat}}$  lớn. U-Net sử dụng một lớp tích chập  $1 \times 1$  để ánh xạ  $C_{\text{feat}}$  kênh này về  $K$  kênh tương ứng với  $K$  lớp cần phân đoạn:

$$\mathbb{R}^{H \times W \times C_{\text{feat}}} \xrightarrow{\text{Conv } 1 \times 1} \mathbb{R}^{H \times W \times K}.$$

Sau đó, áp dụng softmax (đa lớp) hoặc sigmoid (đa nhãm) tại mỗi pixel để thu được phân bố xác suất theo lớp. Các hàm mất mát cụ thể (Dice loss, Dice + Cross-Entropy) được trình bày chi tiết trong phần Hàm mất mát.

**UNet 2D cho phân đoạn u não.** Bài toán này sử dụng UNet 2D, xử lý từng lát cắt 2D được trích từ thể tích MRI 3D của bộ dữ liệu BraTS. Tất cả các phép tích chập, pooling và upsampling đều là 2D. Đầu vào của mạng là các lát cắt kích thước  $H \times W$  với nhiều kênh (các modality MRI khác nhau), đầu ra là bản đồ phân đoạn 2D cùng kích thước, với các kênh tương ứng các vùng khối u quan tâm (WT, TC, ET). Cách tiếp cận này giúp giảm chi phí tính toán so với UNet 3D, nhưng vẫn tận dụng được sức mạnh của kiến trúc chữ U và các skip connection trong việc phân đoạn cấu trúc não bộ phức tạp.

### 3.3 UNet++

UNet++ [?] là một biến thể nâng cấp của UNet, được thiết kế đặc biệt để cải thiện khả năng phân đoạn ảnh y khoa. Hạn chế chính của UNet nằm ở skip connections nguyên bản, vốn nối trực tiếp feature từ encoder sang decoder mặc dù hai loại feature này có mức độ trừu tượng khác nhau.

UNet++ giải quyết vấn đề đó thông qua:

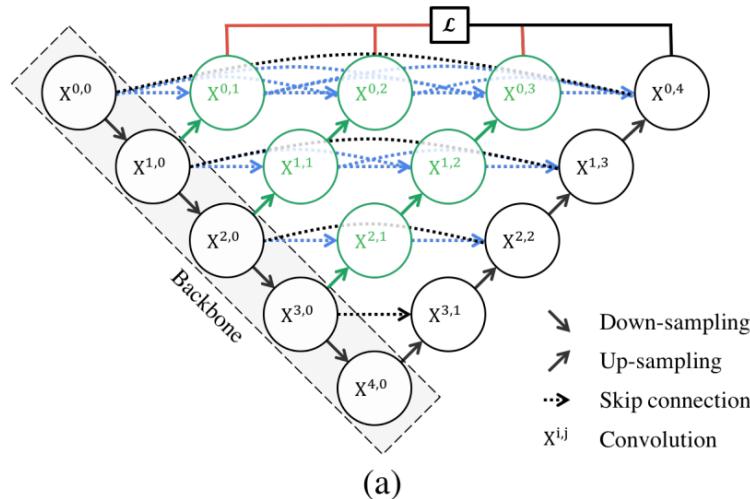
1. Thiết kế lại skip connections
2. Deep supervision
3. Khả năng pruning linh hoạt

Trong UNet:

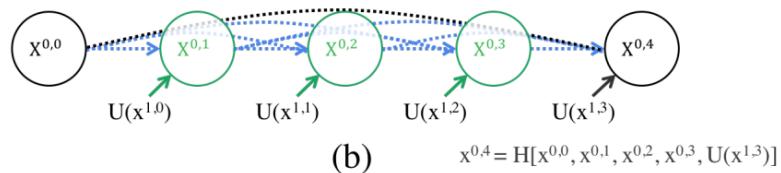
- Feature encoder mang chi tiết không gian, mức trừu tượng thấp.
  - Feature decoder ở giai đoạn sâu lại giàu ngữ nghĩa, mức trừu tượng cao
- Khi ghép 2 loại feature này, xảy ra semantic gap khiến mô hình học khó hơn.

#### 3.3.1 Skip connections được thiết kế lại trong UNet++

UNet++ không nối thẳng encoder với decoder nữa. Thay vào đó mỗi đường skip sẽ trả qua một chuỗi các khối **dense convolution**: nối tất cả các đầu vào trước đó + feature được upsample từ tầng sâu hơn → Nâng dần mức ngữ nghĩa của feature encoder để tiệm cận decoder trước khi ghép vào nhau → mạng học dễ dàng hơn, kết quả phân đoạn chi tiết hơn.



$$x^{0,1} = H[x^{0,0}, U(x^{1,0})] \quad x^{0,2} = H[x^{0,0}, x^{0,1}, U(x^{1,1})] \quad x^{0,3} = H[x^{0,0}, x^{0,1}, x^{0,2}, U(x^{1,2})]$$



Hình 3.2: Kiến trúc tổng quan và skip connections của UNet++

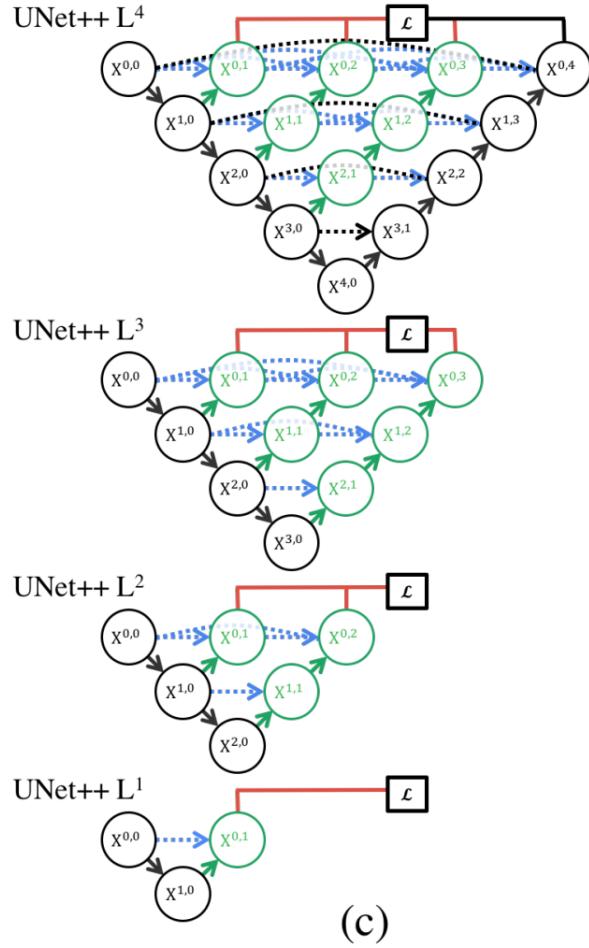
### 3.3.2 Deep Supervision trong UNet++

Mục đích:

- Giúp gradient lan truyền trực tiếp vào các tầng giữa → huấn luyện nhanh và ổn định hơn
- Cho phép lựa chọn nhiều đầu ra phân đoạn tương ứng với các độ sâu khác nhau

2 chế độ:

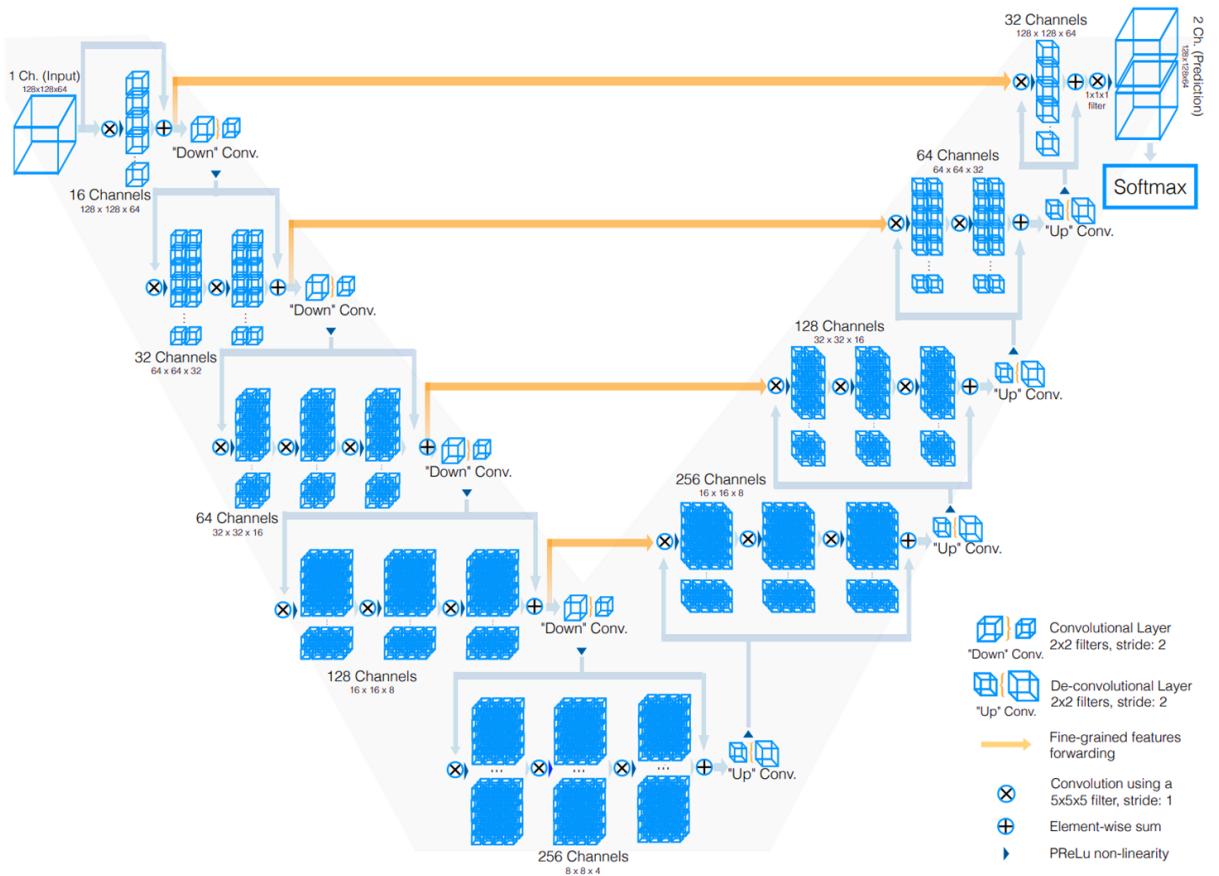
- **Accurate mode:** Dùng tất cả các nhánh phân đoạn, tổng hợp hoặc lấy trung bình đầu ra → kết quả chính xác nhất.
- **Fast mode:**
  - Chỉ chọn một nhánh ( $L^1, L^2$  hoặc  $L^4$ ) để inference
  - Các nhánh còn lại có thể cắt bỏ (prune)
  - Giảm mạnh kích thước mô hình và thời gian chạy inference



### 3.4 V-Net

V-Net [?] được thiết kế như một mạng fully convolutional 3D dành riêng cho các bài toán phân đoạn thể tích trong ảnh y khoa. Khác với các mô hình 2D truyền thống chỉ xử lý từng lát cắt, V-Net học trực tiếp trên toàn bộ thể tích 3D và sinh ra bản đồ phân đoạn 3D cùng kích thước, cho phép mô hình nắm bắt trọng lượng ngữ cảnh không gian theo ba chiều.

Kiến trúc của V-Net gồm hai phần đối xứng tạo thành dạng chữ V: một nhánh nén đặc trưng (compression path) và một nhánh giải nén (decompression path), được kết nối với nhau bằng các skip connection. Sơ đồ tổng quan của mô hình được minh họa trong Hình 3.3.



Hình 3.3: Sơ đồ tổng quan kiến trúc V-Net

## □ Kiến trúc chữ V và cơ chế học đặc trưng

Trong nhánh encoder ở bên trái, V-Net trích xuất đặc trưng ở nhiều mức độ phân giải khác nhau. Mỗi mức (stage), mô hình sử dụng 1 đến 3 lớp tích chập 3D với kernel  $5 \times 5 \times 5$  và padding phù hợp để duy trì kích thước không gian trong phạm vi stage. Các lớp tích chập được kết hợp với hàm kích hoạt Parametric ReLU (PReLU). PReLU là một biến thể của ReLU cho phép học hệ số âm thay vì cố định ở 0.

Hàm PReLU được định nghĩa như sau:

$$f(x) = \begin{cases} x, & x \geq 0, \\ ax, & x < 0. \end{cases}$$

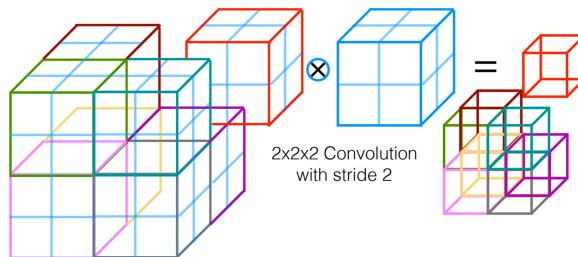
trong đó  $\alpha$  là một tham số có thể học được. Ưu điểm của PReLU là tránh việc triệt tiêu hoàn toàn thông tin tại các vùng đầu vào âm như ReLU truyền thống, từ đó cải thiện khả năng tối ưu và tốc độ hội tụ.

Một đặc điểm nữa của thiết kế V-Net là việc sử dụng residual block cho từng stage. Đầu vào  $x$  được đưa qua chuỗi tích chập – PReLU để tạo tín hiệu biến đổi  $F(x)$ , sau đó được cộng trực tiếp trở lại với chính  $x$ , tạo thành:

$$y = x + F(x)$$

Sự kết hợp dạng residual này giảm thiểu hiện tượng gradient vanishing, cho phép mạng sâu hơn nhưng vẫn ổn định trong quá trình huấn luyện.

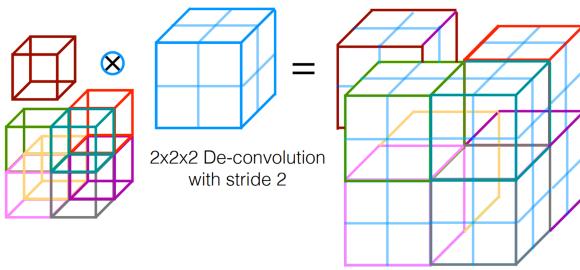
Để giảm kích thước thước không gian khi đi sâu vào mạng, V-Net không dùng max-pooling như U-Net mà thay thế hoàn toàn bằng tích chập 3D có stride = 2 với kernel  $2 \times 2 \times 2$  giúp giảm một nửa kích thước theo mỗi chiều đồng thời tăng gấp đôi số kênh. Điều này không chỉ làm giàu biểu diễn đặc trưng mà còn mang lại ưu điểm so với pooling: mạng có thể học được cách tổng hợp thông tin khi downsample, thay vì dùng quy tắc cứng như max-pooling. Cơ chế downsampling này được mô tả trong Hình 3.4.



Hình 3.4: VNet downsample sử dụng tích chập 3D với kernel size  $2 \times 2 \times 2$  và stride = 2

#### □ Nhánh decoder và skip connections

Nhánh decoder bên phải thực hiện khôi phục lại độ phân giải không gia, đảo ngược quá trình nén đặc trưng. Mỗi stage trong nhánh này bắt đầu bằng một lớp deconvolution (transposed convolution) 3D với kernel  $2 \times 2 \times 2$  và stride = 2. Phép toán này mở rộng kích thước không gian gấp đôi trong khi giảm số lượng kênh đặc trưng, đóng vai trò tương ứng với phép downsample ở nhánh trái. Hình 3.5 minh họa cơ chế upsampling của nhánh decoder.



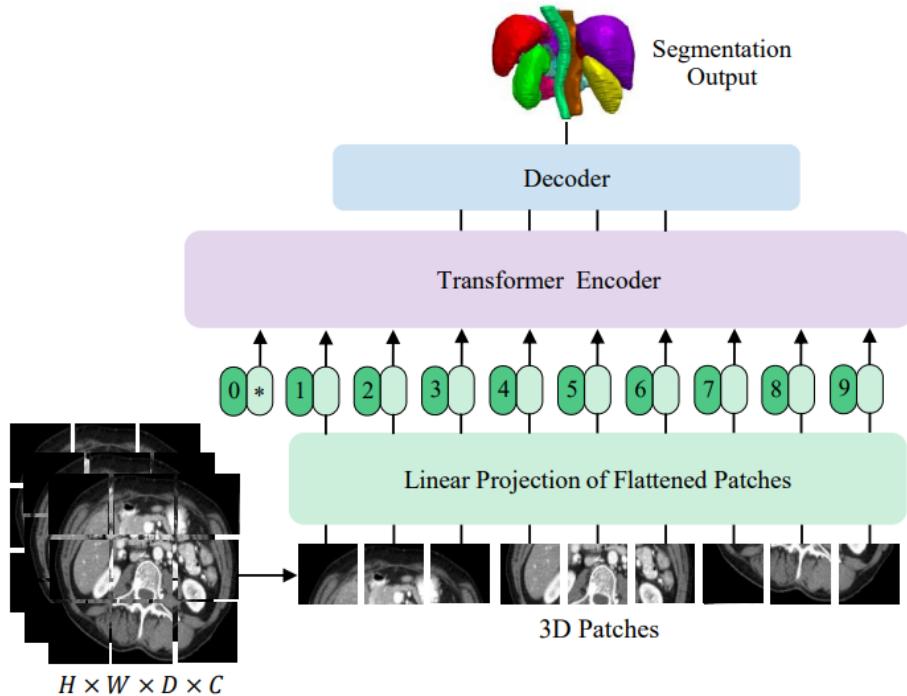
Hình 3.5: VNet upsampling sử dụng transposed convolution 3D với kernel size  $2 \times 2 \times 2$  và stride = 2

V-Net sử dụng các skip connection giữa các stage đối xứng của hai nhánh. Sau mỗi lần upsampling, feature map từ decoder được concatenate với feature map ở cùng độ phân giải từ encoder. Nhờ đó, mô hình kết hợp được cả thông tin chi tiết ở độ phân giải cao (từ encoder) lẫn thông tin trừu tượng ở độ phân giải thấp (từ decoder). Cơ chế này giúp tăng độ chính xác phân đoạn, đặc biệt là tại các vị trí biên mờ hoặc cấu trúc nhỏ.

Sau khi đi qua toàn bộ các stage decoder, VNet sử dụng một lớp tích chập  $1 \times 1 \times 1$  để đưa số kênh về đúng số lớp cần phân đoạn trước khi áp dụng softmax voxel-wise để sinh phân bố xác xuất cho từng voxel.

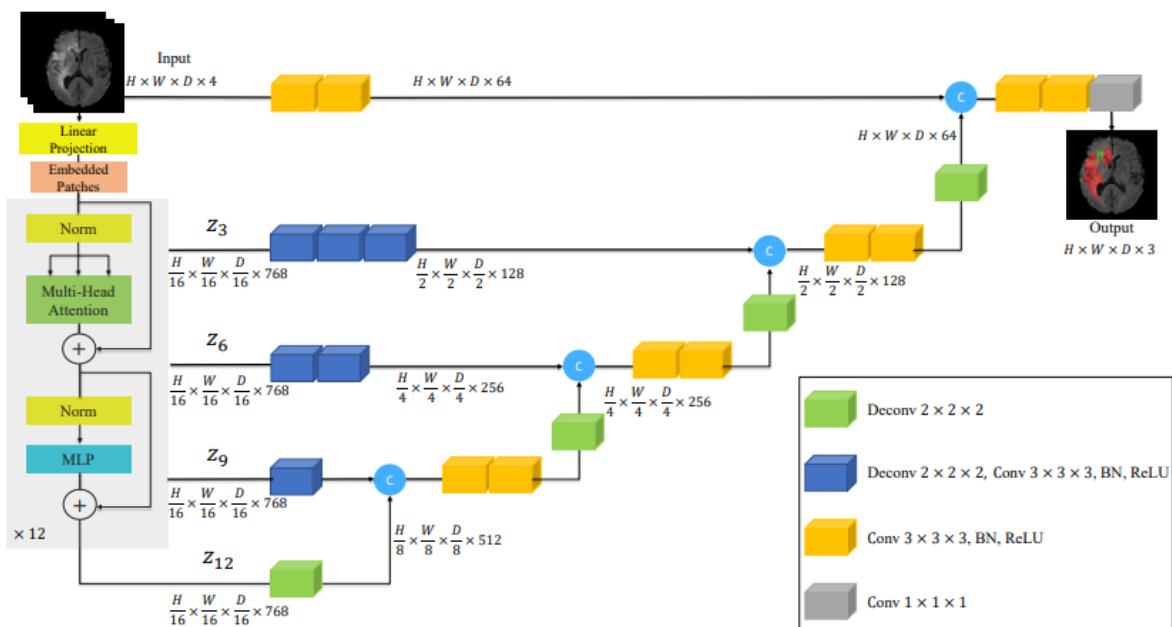
### 3.4.1 UNETR

UNETR (UNet TRansformers) [?] là kiến trúc lai giữa Transformer và CNN dành cho phân đoạn ảnh y tế 3D. Để khắc phục hạn chế về vùng tiếp nhận cục bộ của các mạng tích chập truyền thống, UNETR sử dụng encoder Transformer để mã hóa ảnh dưới dạng chuỗi patch 3D, giúp học hiệu quả các phụ thuộc dài hạn và ngữ cảnh toàn cục. Các biểu diễn này được kết hợp với một decoder CNN thông qua hệ thống skip connection đa độ phân giải, giúp mô hình vừa nắm bắt thông tin diện rộng, vừa duy trì được các chi tiết không gian cục bộ để dự đoán kết quả phân đoạn chính xác.



Hình 3.6: Tổng quan về UNETR

Kiến trúc của UNETR gồm hai phần đối xứng theo dạng chữ U: một nhánh encoder sử dụng transformer để trích xuất đặc trưng toàn cục từ thể tích đầu vào và một nhánh decoder để khôi phục độ phân giải, trong đó hai nhánh được kết nối với nhau thông qua các skip connection ở nhiều mức độ phân giải nhằm tạo ra kết quả phân đoạn ngữ nghĩa cuối cùng. Sơ đồ tổng quan mô hình được minh họa trong các hình dưới đây:



Hình 3.7: Sơ đồ tổng quan kiến trúc UNETR

## □ Biểu diễn đầu vào và embedding

Thể tích đầu vào 3D

$$\mathbf{x} \in \mathbb{R}^{H \times W \times D \times C}$$

được chia thành các patch 3D không chồng lấp có kích thước  $(P, P, P)$ . Các patch này được làm phẳng và sắp xếp thành một chuỗi một chiều, sau đó được chiếu vào không gian embedding  $K$  chiều thông qua một lớp tuyến tính. Để giữ thông tin vị trí không gian, embedding vị trí một chiều có thể học được được cộng vào embedding của mỗi patch. UNETR không sử dụng token [class] do mô hình phục vụ cho bài toán phân đoạn.

## □ Encoder dựa trên Transformer

Encoder bao gồm một chồng các khối transformer, mỗi khối gồm Multi-Head Self-Attention (MSA) và MLP, kết hợp với Layer Normalization và residual connection. Cơ chế self-attention cho phép mô hình học được quan hệ dài hạn giữa các patch trong toàn bộ thể tích 3D.

## □ Trích xuất đặc trưng và skip connection

Tại một số tầng Transformer trung gian, các biểu diễn chuỗi được reshape thành tensor 3D và được chiếu sang không gian đặc trưng thông qua các lớp tích chập  $3 \times 3 \times 3$ . Các đặc trưng này được truyền trực tiếp tới decoder thông qua skip connection, giúp bảo toàn thông tin không gian chi tiết.

## □ Decoder và đầu ra

Decoder thực hiện quá trình upsampling từng bước bằng các lớp deconvolution, bắt đầu từ đầu ra Transformer cuối cùng đóng vai trò nút thắt cổ chai (*bottleneck*) ở độ phân giải thấp nhất. Các đặc trưng này được kết hợp với các đặc trưng từ encoder thông qua skip connection và được tinh chỉnh bằng các lớp tích chập. Cuối cùng, một lớp tích chập  $1 \times 1 \times 1$  kết hợp với hàm softmax tạo ra bản đồ phân đoạn ngữ nghĩa theo từng voxel.

## 3.5 Swin UNet3D

Trong những năm gần đây, các phương pháp phân đoạn ảnh y khoa 3D chủ yếu dựa trên 3 hướng tiếp cận chính:

1. **Các mô hình thuần CNN:** hiệu quả vượt trội trong việc học các đặc trưng cục bộ, tuy nhiên gặp khó khăn trong việc nắm bắt các mối quan hệ phụ thuộc dài hạn và ngữ cảnh toàn cục.

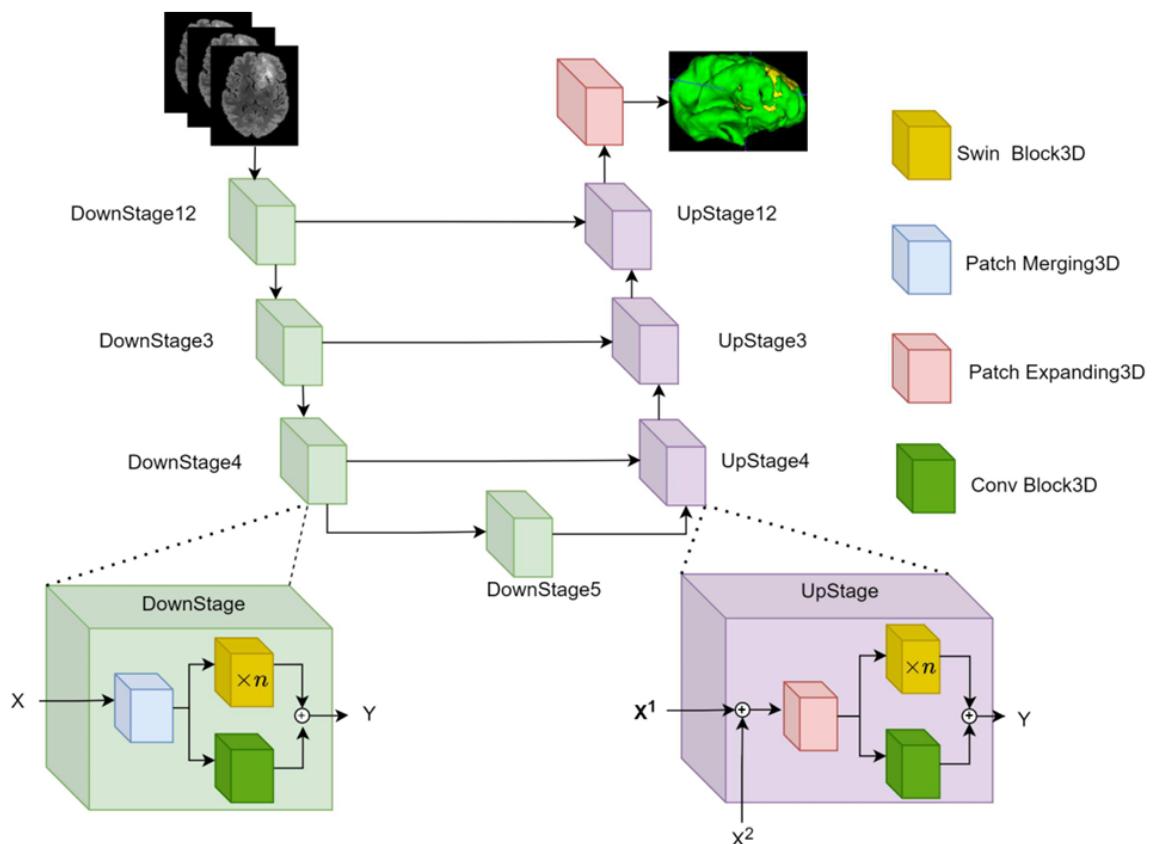
**2. Các mô hình thuần Vision Transformer:** xử lý dữ liệu ảnh dưới dạng token một chiều và sử dụng cơ chế self-attention để mô hình hóa các mối quan hệ phụ thuộc xa. Mặc dù khai thác hiệu quả ngữ cảnh toàn cục, việc thiếu các inductive bias vốn có của tích chập khiến các mô hình ViT gặp phải hạn chế trong việc học các đặc trưng cục bộ và chi tiết hình học quan trọng.

**3. Các kiến trúc lai kết hợp CNN và ViT:** nhằm tận dụng ưu điểm của 2 loại mô hình. Tuy nhiên, phần lớn các mô hình lai dựa trên cơ chế self-attention toàn cục của Transformer, dẫn đến sự gia tăng đáng kể về độ phức tạp tính toán và số lượng tham số.

⇒ Swin UNet3D được đề xuất để giải quyết hạn chế trên, trong đó cơ chế attention theo cửa sổ và dịch cửa sổ được kết hợp với cấu trúc U-Net và các khối tích chập 3D để đạt được sự cân bằng giữa hiệu quả biểu diễn và chi phí tính toán.

### 3.5.1 Tổng quan kiến trúc

Swin UNet3D tuân theo khung UNet: Encoder → Decoder và skip connections theo từng mức phân giải.



Hình 3.8: Kiến trúc tổng thể và các khối chính trong từng stage của Swin UNet3D

Hình 3.8 minh họa kiến trúc tổng thể và các khối chính trong từng stage. Trong kiến trúc:

- **Patch Merging3D**: giảm độ phân giải (downsampling) trong encoder.
- **Patch Expanding3D**: tăng độ phân giải (upsampling) trong decoder.
- **Swin Block3D**: trích xuất đặc trưng dựa trên attention theo cửa sổ; học phụ thuộc xa và thông tin toàn cục trong ảnh
- **Conv Block3D**: trích xuất đặc trưng cục bộ, học phụ thuộc gần.
- Ở mỗi stage, đặc trưng từ nhánh Swin và nhánh Conv được kết hợp để tạo biểu diễn giàu thông tin hơn.

### 3.5.2 Biểu diễn đầu vào theo voxel patch và token

Đầu vào là ảnh 3D

$$X \in \mathbb{R}^{H \times W \times D \times C}$$

Bước đầu tiên là chia ảnh 3D thành các voxel patch không chồng lấp kích thước  $4 \times 4 \times 4$ . Mỗi patch sau đó:

1. Flatten thành vector 1D
2. Qua linear embedding để ánh xạ vào không gian đặc trưng.

Kết quả: ảnh được mã hóa thành tensor token có kích thước

$$\frac{H}{4} \times \frac{W}{4} \times \frac{D}{4} \times C$$

### 3.5.3 Thiết kế theo stage: DownStage và Upstage

Kiến trúc được tổ chức theo các stage:

- **DownStage (Encoder)**: thực hiện giảm kích thước không gian và tăng mức trừu tượng của đặc trưng.

Cấu thành điển hình:

- Patch Merging3D
  - Nhiều Swin Block3D
  - Nhánh Conv Block3D
  - Hợp nhất đặc trưng
  - **UpStage (Decoder)**: khôi phục độ phân giải dần về kích thước đầu vào
- Cấu thành điển hình:

- Patch Expanding3D
- Nhiều Swin Block3D
- Nhánh Conv Block3D
- Kết hợp skip connection từ encoder (cùng mức phân giải)
- Hợp nhất đặc trưng

## 3.6 Hàm mất mát

### 3.6.1 Dice loss cho phân đoạn đa lớp

Dice coefficient là độ đo phổ biến trong phân đoạn y sinh và đặc biệt hiệu quả đối với dữ liệu mất cân bằng. Với  $p_{c,i}$  là xác suất dự đoán voxel  $i$  thuộc lớp  $c$  và  $g_{c,i}$  là nhãn one-hot tương ứng, hệ số Dice cho mỗi lớp được định nghĩa:

$$\text{Dice}_c = \frac{2 \sum_i p_{c,i} g_{c,i} + \epsilon}{\sum_i p_{c,i}^2 + \sum_i g_{c,i}^2 + \epsilon}$$

Dice đa lớp được tính bằng trung bình trên  $C$  lớp:

$$\text{Dice}_{\text{mean}} = \frac{1}{C} \sum_{c=1}^C \text{Dice}_c$$

Dice loss được xác định:

$$\mathcal{L}_{\text{Dice}} = 1 - \text{Dice}_{\text{mean}}$$

**Ưu điểm:**

- Ít nhạy với mất cân bằng lớp.
- Tối ưu trực tiếp độ chồng lấp vùng, phù hợp với chỉ số đánh giá trong BraTS.

**Hạn chế:**

- Gradient có thể kém ổn định ở giai đoạn đầu huấn luyện.

### 3.6.2 Hàm mất mát kết hợp Dice + Cross-Entropy

Cross-entropy (CE) cung cấp tín hiệu gradient ổn định ở mức voxel, trong khi Dice loss tập trung vào tối ưu vùng. Để tận dụng ưu điểm của cả hai, hàm mất mát kết hợp được sử dụng.

Multiclass cross-entropy được định nghĩa:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C g_{c,i} \log p_{c,i}$$

Hàm mất mát DiceCE:

$$\mathcal{L}_{DiceCE} = \mathcal{L}_{Dice} + \mathcal{L}_{CE}$$

**Ưu điểm:**

- Kết hợp tối ưu vùng (Dice) và phân loại voxel ổn định (CE).
- Cải thiện hội tụ và độ ổn định trong huấn luyện so với Dice loss đơn lẻ.

# Chương 4

## Thực nghiệm

### 4.1 Các độ đo đánh giá

Để đánh giá chất lượng mô hình phân đoạn u não trên bộ dữ liệu BraTS2020, chúng ta sử dụng 4 độ đo: Dice Similarity Coefficient (DSC), Intersection over Union (IoU), Average Surface Distance (ASD) và 95th percentile Hausdorff Distance (HD95). Các độ đo này được tính độc lập trên ba vùng giải phẫu:

- Whole Tumor (WT): tập hợp tất cả vùng u, tương ứng với nhãn 1 (NCR/NET) + 2 (ED) + 3 (ET)
- Tumor Core (TC): khối u lõi, gồm 1 (NCR/NET) + 3 (ET).
- Enhancing Tumor (ET): vùng u tăng tín hiệu (nhãn 3).

Việc báo cáo kết quả trên từng vùng cho phép đánh giá chi tiết khả năng mô hình phân biệt cấu trúc u phức tạp.

1. **Dice và IoU đánh giá mức độ trùng khớp giữa vùng dự đoán  $P$  và vùng nhãn chuẩn  $G$ , chúng được định nghĩa như sau:**

$$\text{DSC}(P, G) = \frac{2|P \cap G|}{|P| + |G|}$$
$$\text{IoU}(P, G) = \frac{|P \cap G|}{|P \cup G|}$$

Giá trị Dice và IoU nằm trong  $[0, 1]$ , càng cao biểu thị mức độ trùng khớp càng tốt.

## 2. Average Surface Distance (ASD)

ASD đo khoảng cách trung bình giữa bề mặt vùng dự đoán và bề mặt vùng nhãn thực.

Cho 2 tập điểm biên bề mặt  $S_P$  và  $S_G$ , ASD được định nghĩa:

$$\text{ASD}(P, G) = \frac{1}{|S_P| + |S_G|} \left( \sum_{p \in S_P} d(p, S_G) + \sum_{g \in S_G} d(g, S_P) \right)$$

trong đó  $d(x, S)$  là khoảng cách Euclidean nhỏ nhất từ điểm  $x$  đến tập bề mặt  $S$ . ASD phản ánh độ tròn và độ gần giữa bề mặt dự đoán và thực tế. Giá trị ASD càng nhỏ biểu thị phân đoạn càng chính xác.

## 3. 95th Percentile Hausdorff Distance (HD95)

Khoảng cách Hausdorff cổ điển dễ bị ảnh hưởng bởi các điểm nhiễu (outlier). Do đó, trong thực tế, người ta sử dụng HD95 (giá trị phần trăm thứ 95 của phân bố khoảng cách bề mặt hai phía). Cho 2 bề mặt  $S_P$  và  $S_G$ :

$$\text{HD95}(P, G) = \max\{\text{percentile}_{95}[d(p, S_G)], \text{percentile}_{95}[d(g, S_P)]\}$$

HD95 đánh giá sai lệch biên một cách ổn định và không bị chi phối bởi các điểm bất thường đơn lẻ. Giá trị càng thấp biểu thị sai số biên càng nhỏ.

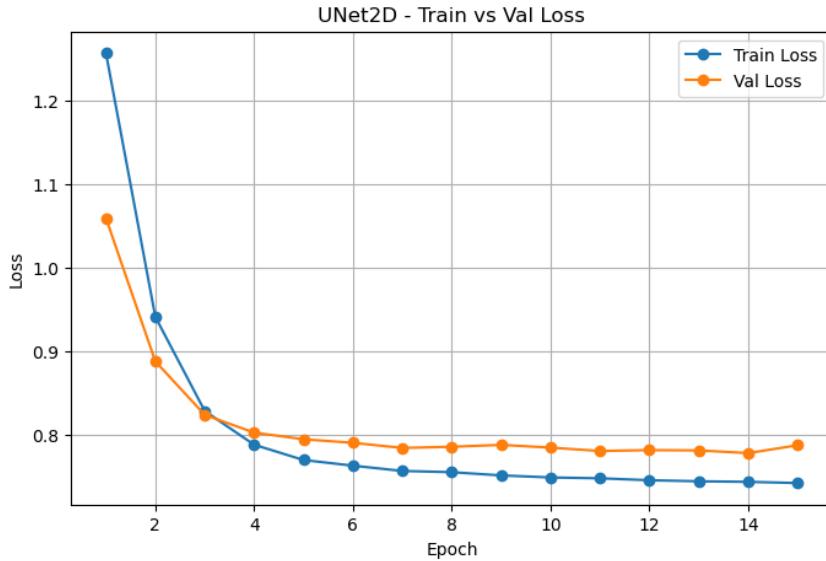
## 4.2 Chi tiết triển khai

### 4.2.1 Mô hình UNet

#### ❑ Thực nghiệm: Huấn luyện mô hình UNet 2D trên các lát cắt 2D từ MRI BraTS2020

- Optimizer Adam với learning rate khởi tạo là  $1e - 3$
- Batch size = 8

Biểu đồ thể hiện quá trình huấn luyện UNet 2D:



Hình 4.1: Quá trình huấn luyện mô hình UNet 2D: loss train và validation

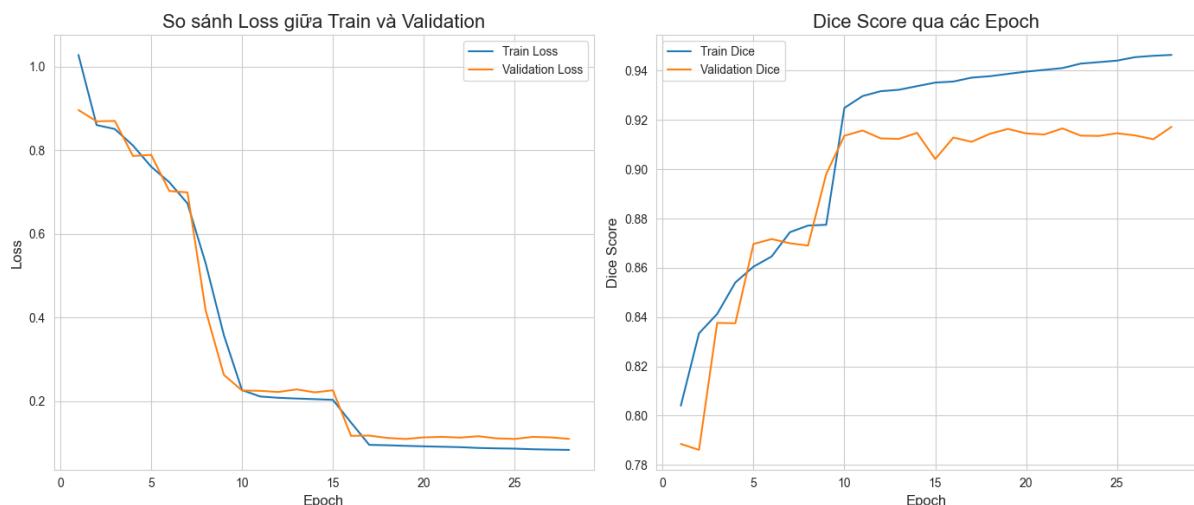
#### 4.2.2 Mô hình UNet++

##### ❑ Thực nghiệm : Huấn luyện mô hình U-Net++ trên các lát cắt 2D từ MRI 3D BraTS2020

Cài đặt huấn luyện:

- Mô hình được huấn luyện với optimizer Adam, batch size = 8, learning rate ban đầu=  $1e - 3$  trong 20 epoch.
- Hàm mất mát DiceCE được sử dụng để huấn luyện mô hình phân đoạn đa nhãn, tối ưu hóa sự trùng khớp giữa dự đoán và ground truth cho từng nhãn riêng biệt.

Biểu đồ thể hiện quá trình training:



Hình 4.2: Biểu đồ thể hiện quá trình training với Unet++

### 4.2.3 Mô hình VNet

□ **Thử nghiệm 1: Huấn luyện mô hình VNet trực tiếp với trên toàn bộ thể tích MRI 3D của từng ca bệnh, không cắt thành patch:**

#### 1. Dữ liệu đầu vào:

- Mỗi ca gồm 4 chuỗi MRI được resize về  $128 \times 128 \times 128$  rồi xếp chồng thành một tensor có kích thước  $(4, 128, 128, 128)$ .
- Ở tập train, áp dụng các phép augmentation 3D: lật ngẫu nhiên theo trực, biến thiên cường độ, nhiễu Gaussian, phóng/thu thể tích,... nhằm tăng đa dạng dữ liệu.

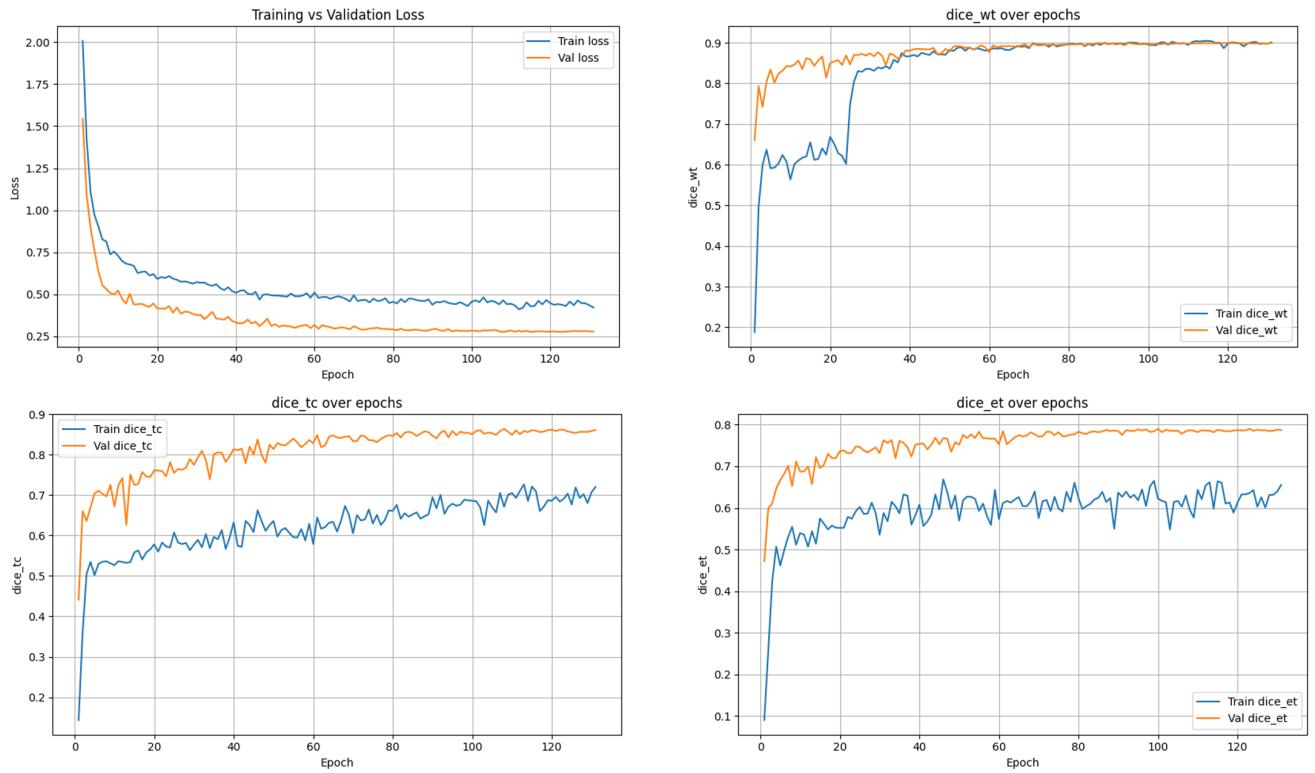
#### 2. Thiết lập huấn luyện:

- Mô hình được huấn luyện với optimizer AdamW, batch size = 2, learning rate ban đầu  $1e - 3$  trong tối đa 200 epoch.
- Hàm mất mát sử dụng: DiceCE, trong đó Dice loss được tính trên các lớp foreground.
- Chiến lược giảm learning rate được áp dụng dựa trên Dice trung bình của ba vùng WT, TC, ET trên tập validation.

#### 3. Quá trình inference và đánh giá:

Ở giai đoạn inference, volume 4 kênh được resize về  $128^3$ , đưa qua mạng để thu được mask dự đoán, sau đó resize mask về kích thước ban đầu để đánh giá.

Biểu đồ thể hiện quá trình training:



Hình 4.3: Thử nghiệm 1 VNET: Quá trình training

## □ **Thử nghiệm 2: Huấn luyện VNet trên các patch 3D thay vì sử dụng toàn bộ thể tích**

### 1. **Dữ liệu đầu vào và chiến lược lấy mẫu patch**

Khác với thử nghiệm đầu tiên, mô hình không dùng toàn bộ volume mà lấy ra patch 3D có kích thước  $128 \times 128 \times 128$  từ khối ảnh. 4 chiến lược lấy mẫu patch bao gồm:

- random → Crop ngẫu nhiên trong toàn thể tích
- rejection sampling → Crop ngẫu nhiên nhưng loại bỏ các patch chứa quá ít foreground
- center\_fg → crop patch được căn giữa quanh một voxel thuộc vùng khồi u, giúp tập trung vào khu vực quan trọng
- mixed → Kết hợp nhiều chiến lược trên theo trọng số

*Trong các thử nghiệm sử dụng chiến lược này với trọng số 0.5 random và 0.5 center\_fg nhằm tăng tính đa dạng và ổn định của các patch*

Cũng sử dụng data augmentation như trong thử nghiệm 1.

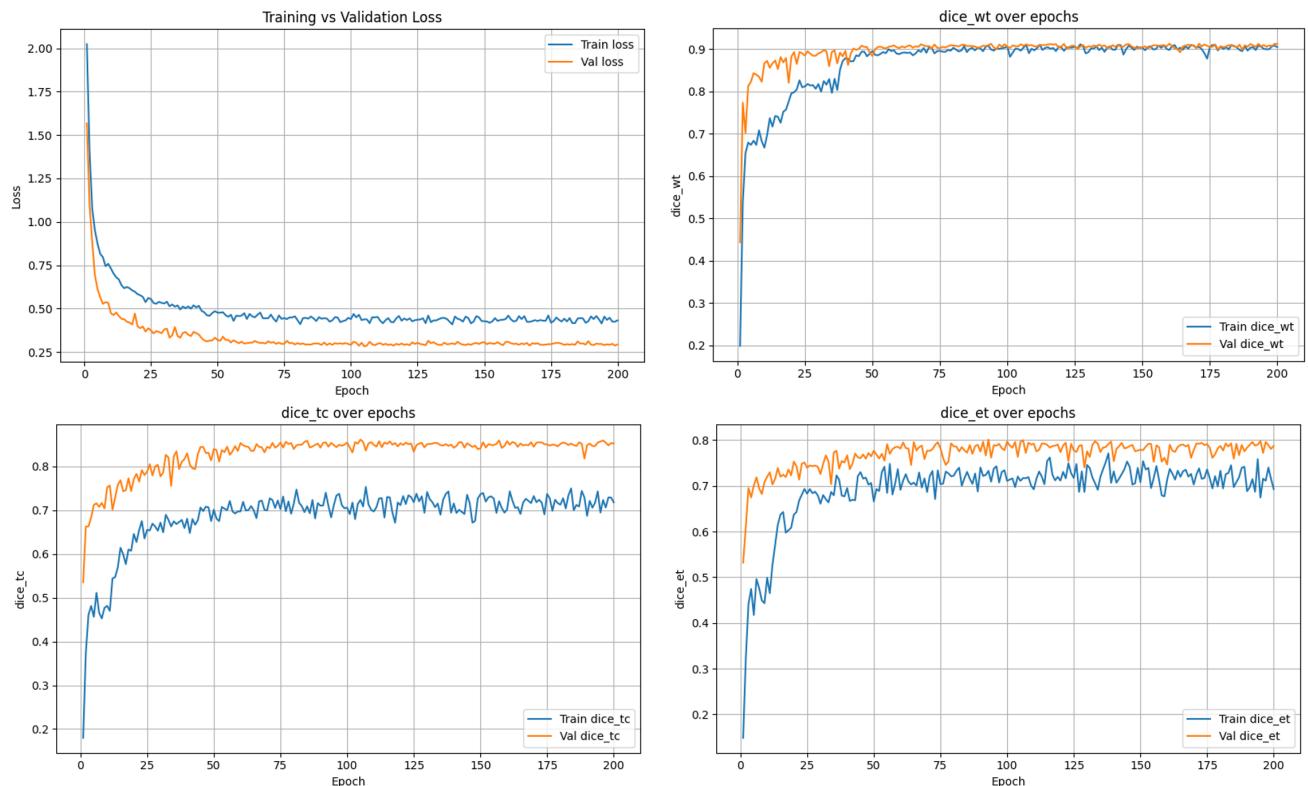
### 2. **Thiết lập huấn luyện tương tự thử nghiệm 1.**

### 3. Quá trình inference và đánh giá:

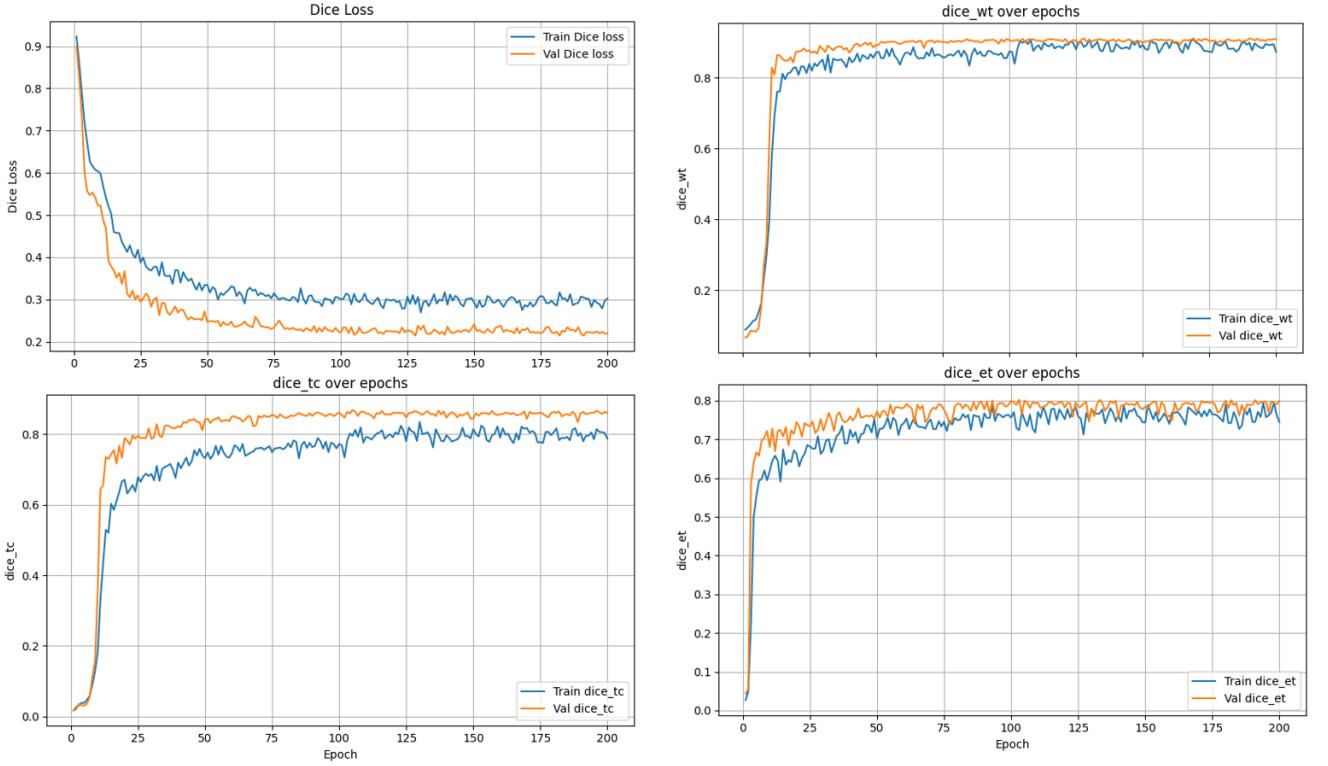
Do mô hình được huấn luyện theo hướng patch-based, giai đoạn inference sử dụng sliding window 3D trên toàn thể tích để tạo ra dự đoán cuối cùng:

- Patch:  $128^3$
- Stride:  $20^3$
- Các vùng chồng lấp được hợp nhất bằng trung bình softmax (probability averaging).
- Dự đoán cuối cùng lấy argmax theo kênh.

Quy trình này đảm bảo tính toàn cục của kết quả phân đoạn dù mô hình được huấn luyện trên patch.



Hình 4.4: Huấn luyện VNet trên các patch 3D sử dụng hàm mất mát hợp DiceLoss + CELoss



Hình 4.5: Huấn luyện VNet trên các patch 3D sử dụng hàm mất mát DiceLoss

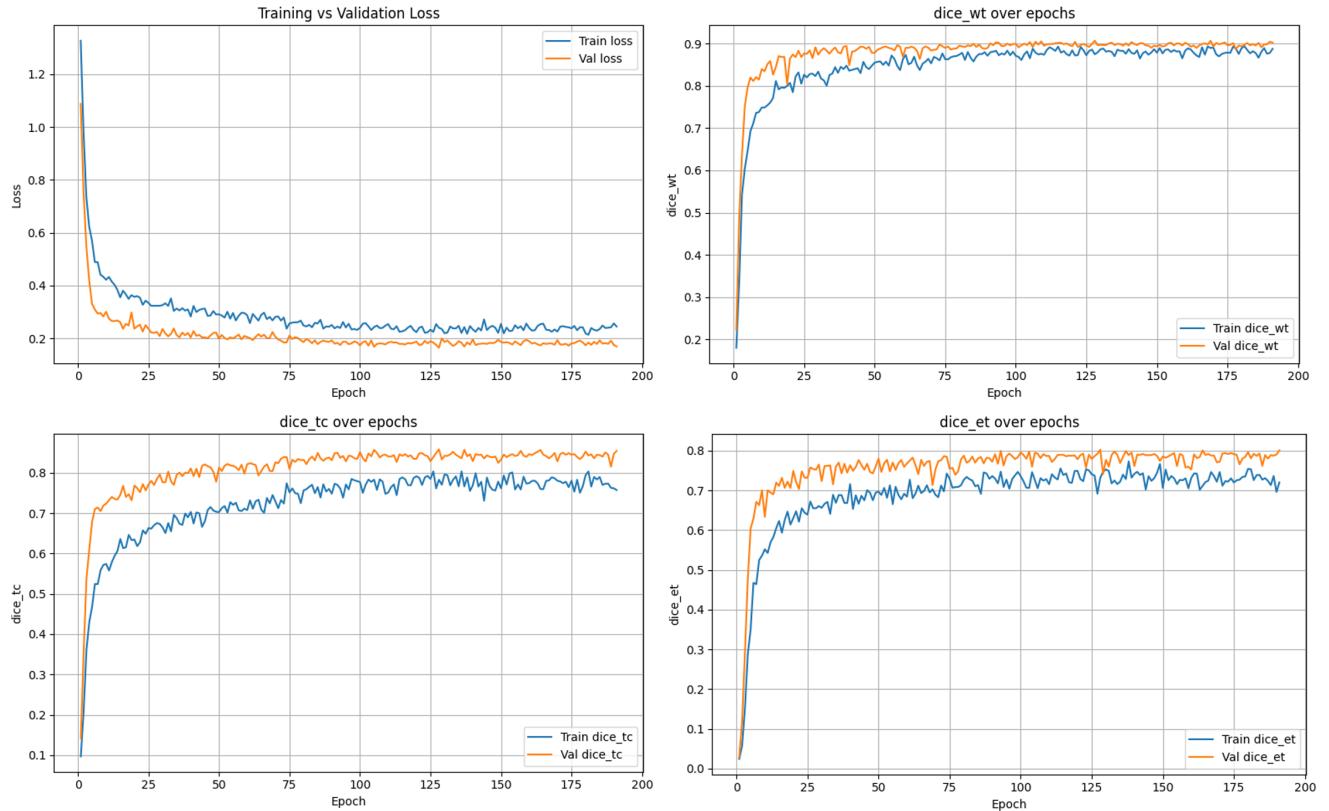
### □ Thử nghiệm 3: Huấn luyện VNet Multi-Head trên các patch 3D

Mở rộng VNet thành VNet Multi-Head. Khác với mô hình VNet tiêu chuẩn sử dụng một nhánh dự đoán duy nhất cho phân đoạn đa lớp, mô hình Multi-Head được thiết kế với một encoder-decoder chung nhưng ba đầu ra (head) độc lập, mỗi head đảm nhận một nhiệm vụ phân đoạn nhị phân riêng:

- WT-head (Whole Tumor): nhận diện toàn bộ vùng u ( $\text{nh}_{\text{an}} > 0$ )
- TC-head (Tumor Core): nhận diện các voxel thuộc nhân u hoặc vùng tăng tín hiệu ( $\text{nh}_{\text{an}} \in \{1, 3\}$ )
- ET-head (Enhancing Tumor): nhận diện vùng u tăng tín hiệu ( $\text{nh}_{\text{an}} = 3$ ).

Cách tiếp cận này được kỳ vọng cho phép mô hình tập trung học đặc trưng cho từng cấu trúc giải phẫu mà không phải cạnh tranh trực tiếp trong một không gian phân lớp chung. Đồng thời, việc chia nhỏ nhiệm vụ thành ba bài toán nhị phân giúp giảm độ phức tạp của hàm mất mát và hạn chế ảnh hưởng bất cân bằng giữa các lớp (đặc biệt ET có số lượng rất nhỏ).

Tương tự các thử nghiệm trước, việc huấn luyện được thực hiện theo chiến lược patch-based 3D nhằm giảm chi phí bộ nhớ và tăng tính đa dạng của mẫu huấn luyện. Mỗi patch được gán ba mặt nạ nhị phân WT, TC và ET tương ứng để huấn luyện từng head. Sự dụng hàm kết hợp giữa DiceLoss nhị phân + CELoss làm hàm mất mát cho mỗi head. Giá trị hàm mất mát cuối cùng là trung bình của 3 head.



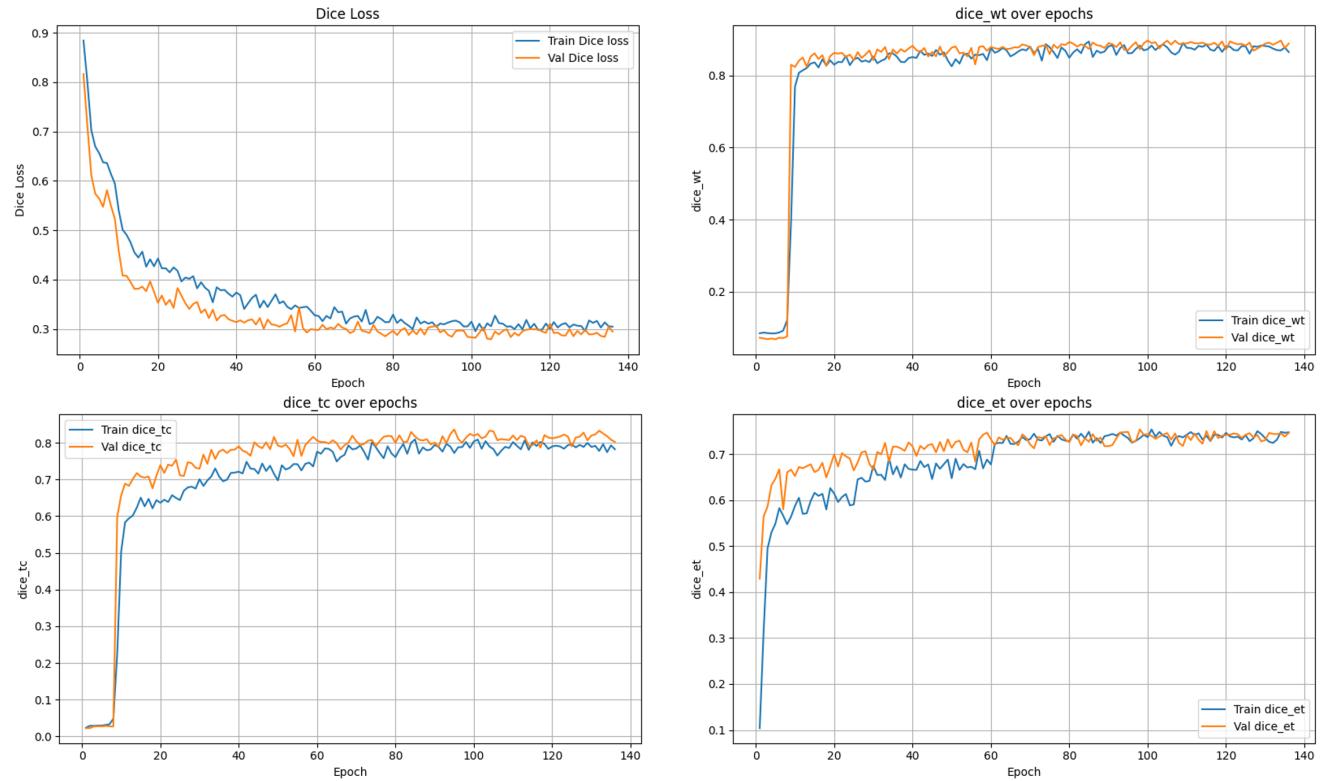
Hình 4.6: Huấn luyện VNet Multi-Head trên các patch 3D

#### □ **Thử nghiệm 4: Huấn luyện VNet Multi-encoder trên các patch 3D**

Khảo sát một biến thể kiến trúc khác của VNet là VNet Multi-encoder. Thay vì xếp chồng 4 kênh đầu vào và đưa qua một encoder duy nhất như các thử nghiệm trước, mô hình VNet multi-encoder sử dụng 4 encoder riêng biệt, mỗi encoder chỉ nhận một modality và học đặc trưng chuyên biệt cho modality đó. Tại mỗi mức độ phân giải trong encoder (tương ứng với các mức 1,...,5 của VNet), các đặc trưng từ 4 nhánh được nối theo chiều kênh và nén lại bằng các lớp tích chập 3D  $1 \times 1 \times 1$  để đưa số kênh về đúng cấu hình VNet gốc. Phần decoder phía sau được giữ nguyên như kiến trúc VNet truyền thống và hoạt động trên các đặc trưng đã được fusion.

Thiết kế này cho phép mô hình tách bạch quá trình trích xuất đặc trưng theo từng modality và quá trình hợp nhất thông tin liên-modality ở mức sâu, thay vì buộc mạng phải học đồng thời

cả hai nhiệm vụ ngay từ các lớp đầu tiên. Về mặt trực giác, các encoder riêng có thể tập trung học những mẫu hình đặc trưng của từng loại ảnh (ví dụ FLAIR nhạy với edema, T1ce nhẫn mạnh vùng u tăng tương phản), trong khi các lớp nén  $1 \times 1 \times 1$ . Đổi lại, mô hình trở nên nặng hơn đáng kể (xấp xỉ gấp bốn lần số tham số ở phần encoder) và yêu cầu bộ nhớ GPU lớn hơn; do đó trong thực nghiệm chúng tôi giảm kích thước batch size = 1 để đảm bảo huấn luyện ổn định trên cùng cấu hình phần cứng.



Hình 4.7: Huấn luyện VNet Multi-Encoder trên các patch 3D

#### 4.2.4 Mô hình UNETR

**Thử nghiệm: Huấn luyện mô hình UNETR trực tiếp trên toàn bộ thể tích MRI 3D của từng ca bệnh, không cắt thành patch:**

##### 1. Dữ liệu đầu vào:

- Mỗi ca gồm 4 chuỗi MRI được resize về  $128 \times 128 \times 128$  rồi xếp chồng thành một tensor có kích thước  $(4, 128, 128, 128)$ .
- Dữ liệu được chuẩn hóa theo z-score trên vùng khác không (*non-zero*), sau đó được cắt ngưỡng (*clipped*) về khoảng  $[-5, 5]$  nhằm hạn chế ảnh hưởng của các giá trị ngoại lai (*outliers*).

- Ở tập train, áp dụng các phép augmentation 3D: lật ngẫu nhiên theo trục, biến thiên cường độ, nhiễu Gaussian, phóng/thu thể tích,... nhằm tăng đa dạng dữ liệu.

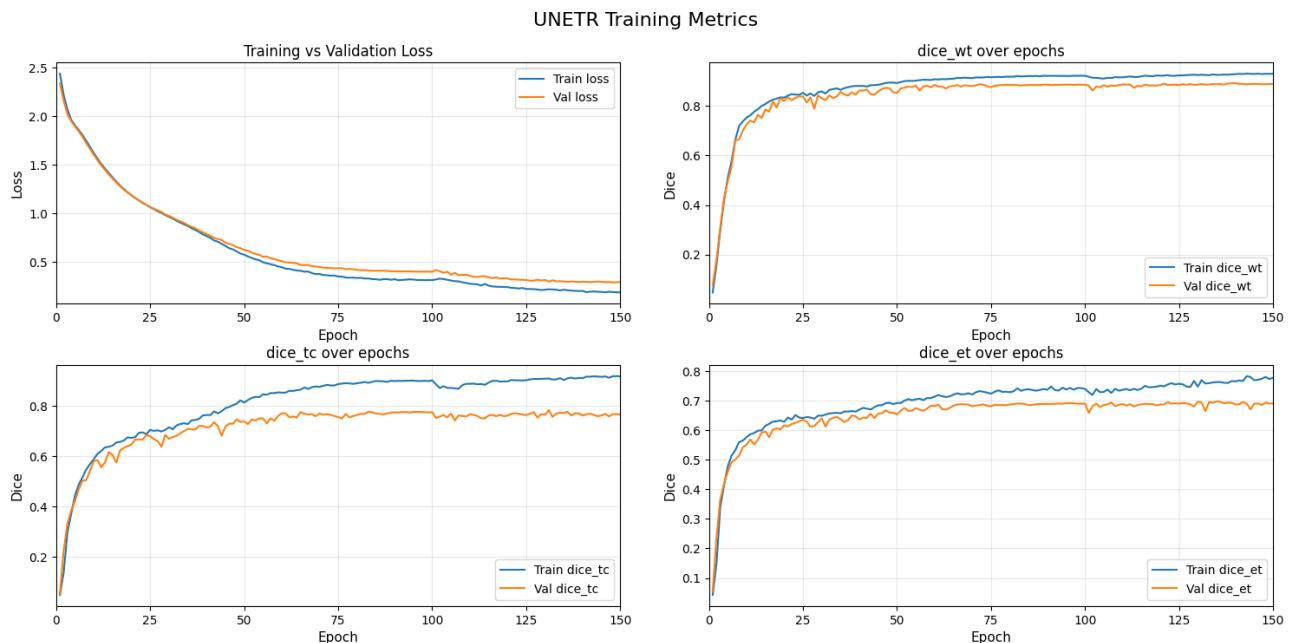
## 2. Thiết lập huấn luyện:

- Mô hình được huấn luyện với optimizer AdamW, batch size = 4, learning rate ban đầu  $1 \times 10^{-4}$  trong tối đa 150 epoch.
- Hàm mất mát sử dụng: là hàm kết hợp (*Combined Loss*) giữa Dice Loss và Cross-Entropy Loss, trong đó Dice Loss được tính trên các lớp foreground (bỏ qua lớp background), còn Cross-Entropy Loss được tính trên toàn bộ bốn lớp.

## 3. Quá trình inference và đánh giá:

Ở giai đoạn inference, volume 4 kênh được resize về  $128^3$ , đưa qua mạng để thu được mask dự đoán, sau đó resize mask về kích thước ban đầu để đánh giá.

Biểu đồ thể hiện quá trình training:



Hình 4.8: Biểu đồ quá trình training với UNETR

### 4.2.5 Mô hình Swin UNet3D

Tương tự thực hiện huấn luyện mô hình Swin UNet 3D trên các patch 3D.

# Chương 5

## Kết quả và thảo luận

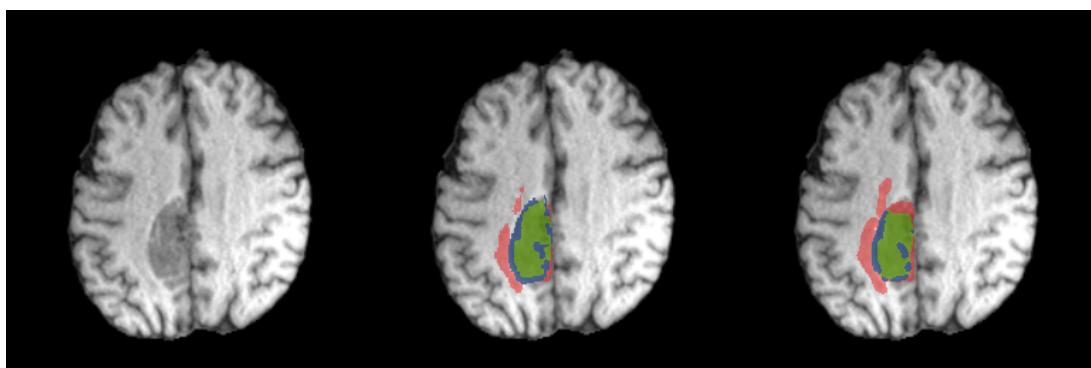
### 5.1 Kết quả

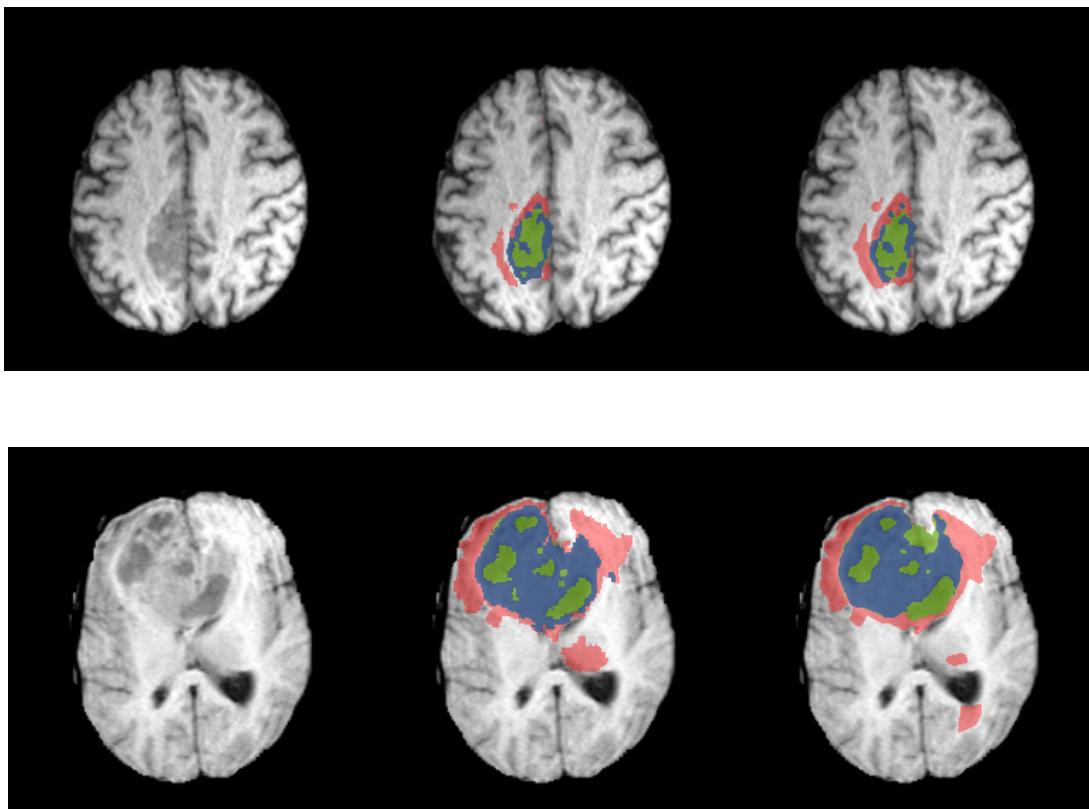
#### □ Mô hình UNet

Đối với các mô hình 2D, quá trình suy luận được thực hiện trên từng lát cắt 2D độc lập. Các kết quả phân đoạn trên từng lát cắt sau đó được xếp chồng lại để tái tạo thành ảnh phân đoạn 3D hoàn chỉnh, dựa trên thông tin spacing của dữ liệu đầu vào. Trên cơ sở ảnh phân đoạn 3D thu được, các chỉ số đánh giá như Dice, IoU, ASD và HD95 được tính toán nhằm đảm bảo tính nhất quán và công bằng khi so sánh với các mô hình 3D.

ROI	Dice	IoU	ASD (mm)	HD95 (mm)
WT	<b>0.8612</b>	0.7725	3.6368	19.0979
TC	<b>0.7549</b>	0.6570	4.6777	12.5280
ET	<b>0.7148</b>	0.5992	9.1472	18.2997

Bảng 5.1: Kết quả inference mô hình UNet trên tập test





Hình 5.1: Ví dụ kết quả phân đoạn của UNet 2D trên một lát cắt: ảnh MRI đầu vào, mask ground truth và mask dự đoán cho các vùng WT, TC, ET

#### □ Mô hình UNet++

<b>Region</b>	<b>Dice</b>	<b>IoU</b>	<b>ASD</b>	<b>HD95</b>
WT	0.895382	0.815754	1.572549	5.021823
TC	0.819587	0.729654	2.006356	6.242132
ET	0.736035	0.627760	1.691734	5.007932

Bảng 5.2: Kết quả inference UNet++ trên tập test

- Đánh giá kết quả mô hình theo các vùng (Region):

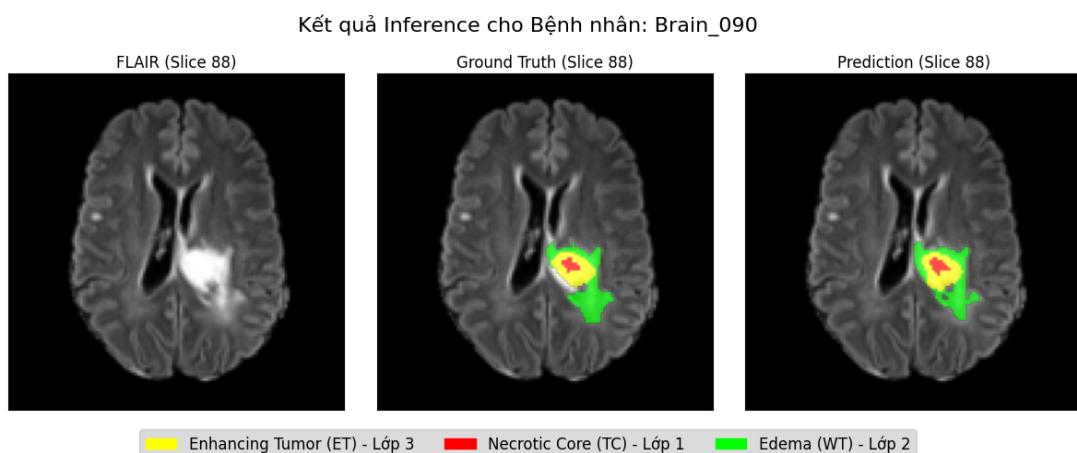
- **Vùng WT (Whole Tumor):**

- Dice = 0.8954, IoU = 0.8158: mô hình đạt hiệu quả cao trong việc phân đoạn toàn bộ khối u, gần như khớp tốt với ground truth.
- ASD = 1.5725, HD95 = 5.0218: sai số khoảng cách trung bình và sai số cực đại 95% đều ở mức chấp nhận được, minh chứng cho độ chính xác không gian khá tốt.

- **Vùng TC (Tumor Core):**

- Dice = 0.8196, IoU = 0.7297: thấp hơn WT, cho thấy việc phân đoạn phần lõi khối u phức tạp hơn nhưng vẫn đạt mức khá tốt.
  - ASD = 2.0064, HD95 = 6.2421: giá trị tăng nhẹ so với WT, phản ánh một số sai lệch trong ranh giới phân đoạn, đặc biệt ở các vùng mép lõi.
- **Vùng ET (Enhancing Tumor):**
- Dice = 0.7360, IoU = 0.6278: thấp nhất trong ba vùng, thể hiện việc dự đoán vùng ET khó khăn nhất, có thể do kích thước nhỏ, hình dạng phức tạp hoặc độ tương phản kém.
  - ASD = 1.6917, HD95 = 5.0079: mặc dù sai số trung bình không quá lớn nhưng HD95 vẫn khá cao, chứng tỏ mô hình gặp khó khăn với một số điểm cực đoan trong vùng ET.

**Nhận xét tổng quan:** Mô hình đạt hiệu quả tốt nhất trên vùng WT, trung bình khá trên vùng TC, và gặp nhiều khó khăn hơn với vùng ET. Các chỉ số ASD và HD95 cho thấy mô hình phân đoạn khá ổn về mặt hình học, nhưng vẫn có sai số ở các chi tiết nhỏ hoặc ranh giới phức tạp. Đây là xu hướng thường gặp trong phân đoạn khối u não, khi các vùng nhỏ và phức tạp như ET thường khó dự đoán chính xác hơn các vùng lớn như WT.



Hình 5.2: Kết quả dự đoán của mô hình trên lát cắt 88 của ca chụp 90

### Nhận xét:

Mô hình thể hiện khả năng phân đoạn tốt các vùng chính, đặc biệt là WT và TC, với hình dạng và vị trí gần như trùng khớp Ground Truth. Vùng ET nhỏ và chi tiết phức tạp nên dự đoán có sai số nhẹ, phản ánh xu hướng khó khăn phổ biến trong phân đoạn tự động các vùng khối

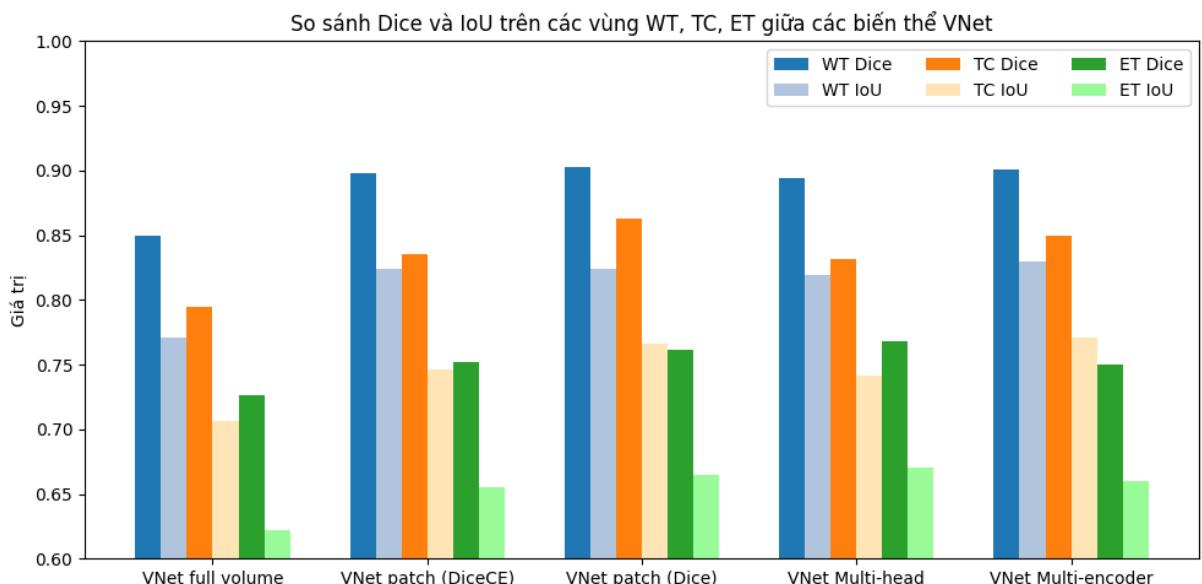
u não nhỏ. Kết quả này cho thấy mô hình có độ chính xác không gian cao và khả năng định vị chính xác các cấu trúc khối u trên lát cắt 2D FLAIR.

## □ Mô hình VNet

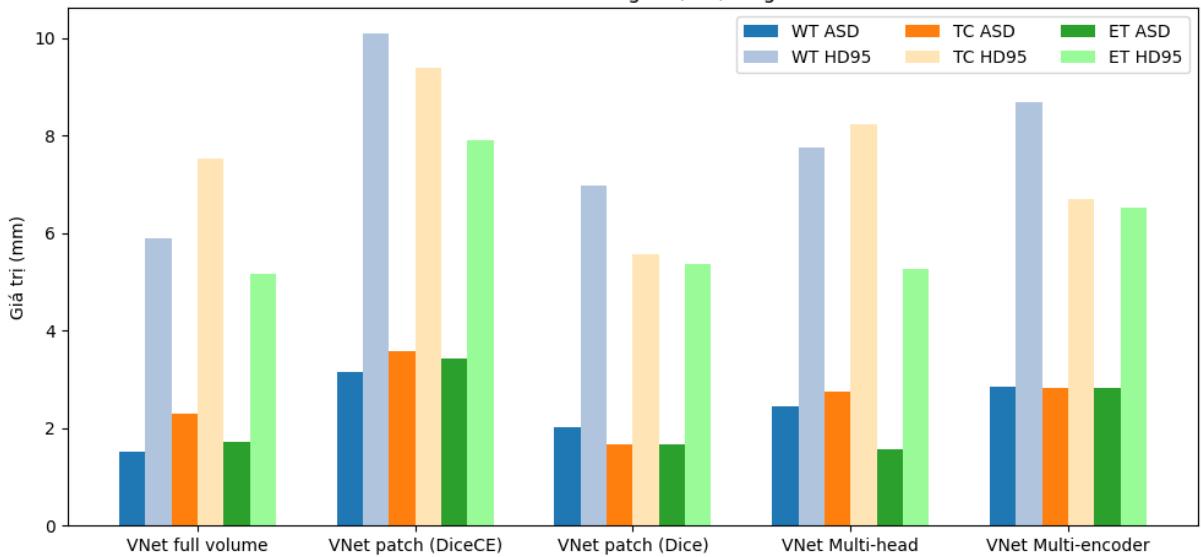
	WT				TC				ET			
	Dice	IoU	ASD	HD95	Dice	IoU	ASD	HD95	Dice	IoU	ASD	HD95
VNet trên toàn bộ thể tích	0.8501	0.7706	1.5178	5.8878	0.7949	0.7068	2.2944	7.5172	0.7265	0.6219	1.7098	5.1759
VNet trên các patch 3D (DiceCELoss)	0.8977	0.8236	3.1626	10.1072	0.8353	0.7464	3.5748	9.3785	0.7519	0.6556	3.4380	7.8940
VNet trên các patch 3D (DiceLoss)	0.9024	0.8236	2.0219	6.9727	0.8633	0.7664	1.6688	5.5709	0.7617	0.6650	1.6565	5.3642
VNet Multi-Head trên các patch 3D	0.8944	0.8195	2.4352	7.7570	0.8315	0.7416	2.7558	8.2220	0.7680	0.6703	1.5727	5.2660
VNet Multi-encoder trên các khối 3D	0.9011	0.8300	2.8607	8.6980	0.8499	0.7706	2.8147	6.6950	0.7498	0.6599	2.8145	6.5197

Bảng 5.3: Kết quả đánh giá các mô hình VNet trên các vùng WT, TC và ET.

Bảng 5.3 trình bày chi tiết các chỉ số đánh giá trên tập test cho toàn bộ các biến thể của VNet, bao gồm huấn luyện trên toàn bộ thể tích, huấn luyện theo patch 3D với hai lựa chọn hàm mất mát (DiceCE và Dice), cũng như hai biến thể mở rộng là VNet multi-head và VNet multi-encoder.

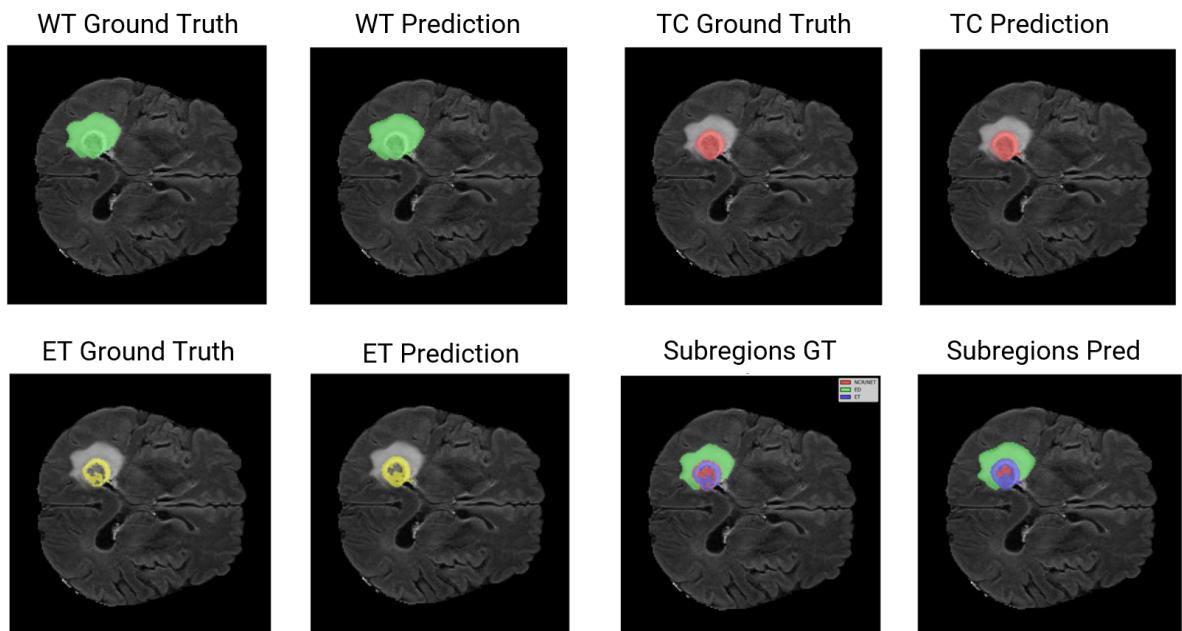


So sánh ASD và HD95 trên các vùng WT, TC, ET giữa các biến thể VNet

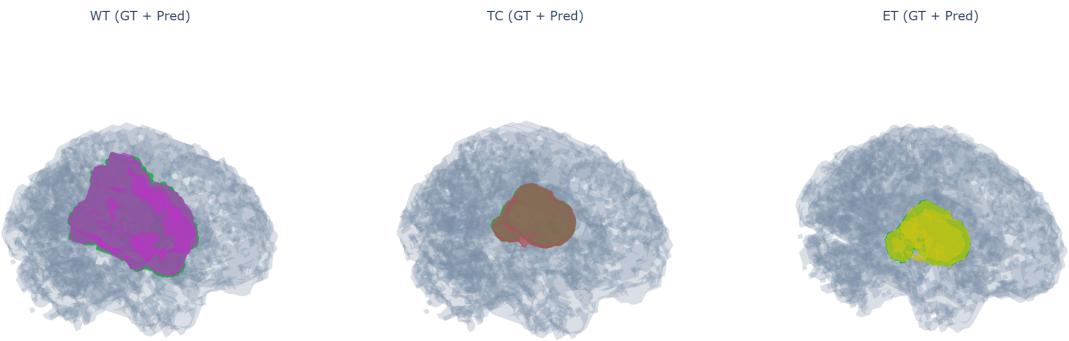


Nhìn chung, trong nhóm các biến thể VNet, huấn luyện trên các patch 3D với DiceLoss thuận tiện cho kết quả tốt nhất xét trung bình ba cấu trúc, trong khi VNet multi-head tỏ ra đặc biệt hiệu quả đối với vùng ET và VNet multi-encoder cải thiện IoU cho WT và TC nhờ khả năng khai thác thông tin từng modality một cách chuyên biệt.

Minh họa kết quả dự đoán của mô hình trên lát cắt  $z = 75$  của ca chụp 11 được overlay lên chuỗi Flair:

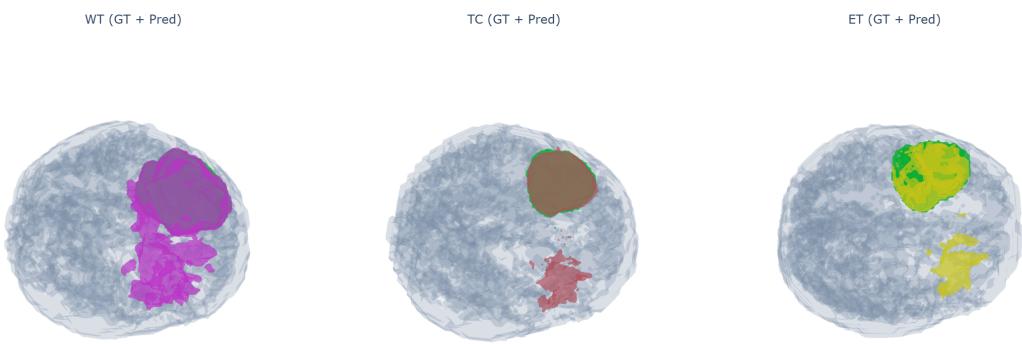


Hình 5.3: Kết quả phân đoạn của VNet trên lát cắt thứ 75 của ca chụp 11



Hình 5.4: Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 104

Minh họa trường hợp mô hình dự đoán tê:



Hình 5.5: Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 193

## □ Mô hình UNETR

<b>Region</b>	<b>Dice</b>	<b>IoU</b>	<b>ASD</b>	<b>HD95</b>
WT	0.8729	0.7887	3.4713	10.5030
TC	0.7819	0.6777	5.2696	12.7303
ET	0.6869	0.5788	4.2527	12.4837

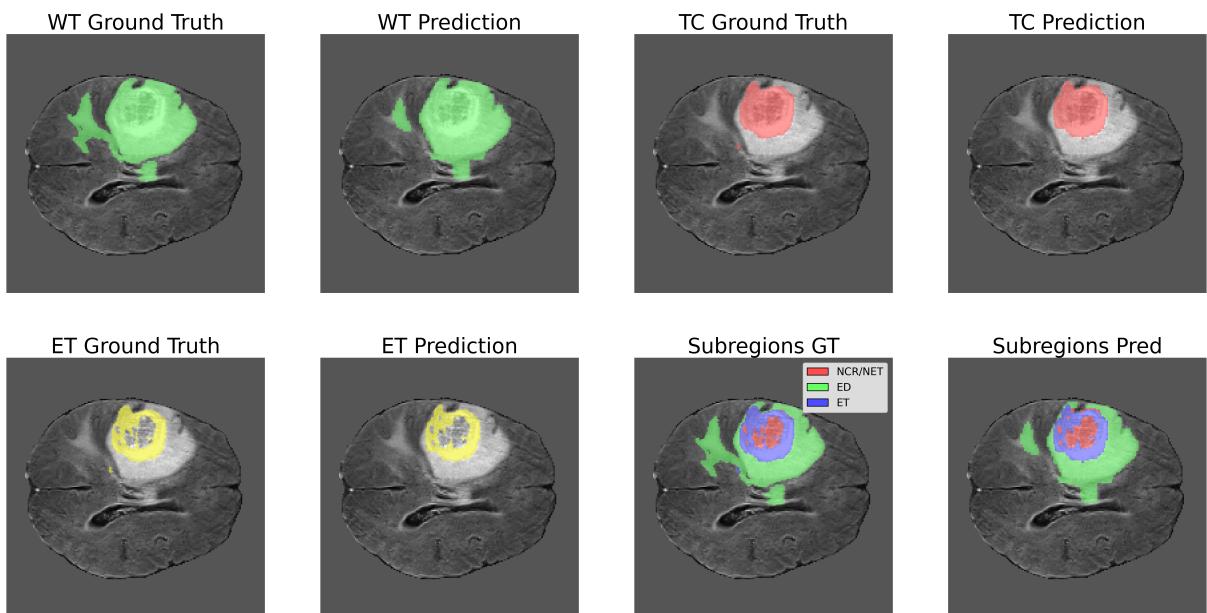
Bảng 5.4: Kết quả đánh giá các mô hình UNETR trên các vùng WT, TC và ET

- Đánh giá kết quả mô hình:

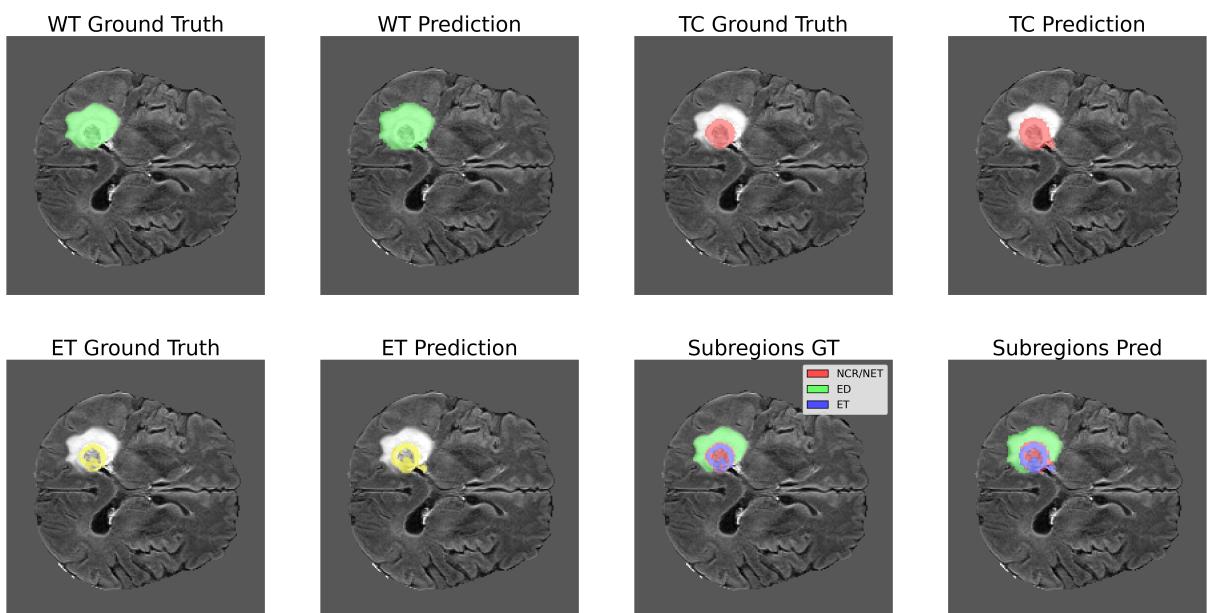
- Mô hình đạt hiệu suất tốt nhất trên vùng WT với Dice = 0.8729, cho thấy khả năng phân đoạn toàn bộ khối u một cách chính xác.
- Vùng TC đạt Dice = 0.7819, phản ánh mức độ khó trong việc xác định ranh giới của lõi u.
- Vùng ET có giá trị Dice thấp nhất (0.6869) do đây là vùng nhỏ nhất và khó phân đoạn nhất, một hiện tượng thường gặp trong các mô hình phân đoạn u não.

- Các chỉ số ASD và HD95 cho thấy sai số khoảng cách bề mặt nằm trong phạm vi chấp nhận được, trong đó vùng WT đạt độ chính xác cao nhất với  $ASD = 3.47$  mm.

Mô hình UNETR đạt hiệu suất khá tốt với Dice trung bình 0.7806 trên ba vùng WT/TC/ET. Điểm mạnh nổi bật là phân đoạn Whole Tumor (Dice = 0.8729) nhờ khả năng học đặc trưng toàn cục của Vision Transformer. Tuy nhiên, vùng Enhancing Tumor còn hạn chế (Dice = 0.6869) do kích thước nhỏ và class imbalance.



Hình 5.6: Kết quả phân đoạn của UNETR trên lát cắt thứ 88 của ca chụp 091



Hình 5.7: Kết quả phân đoạn của UNETR trên lát cắt thứ 75 của ca chụp 011

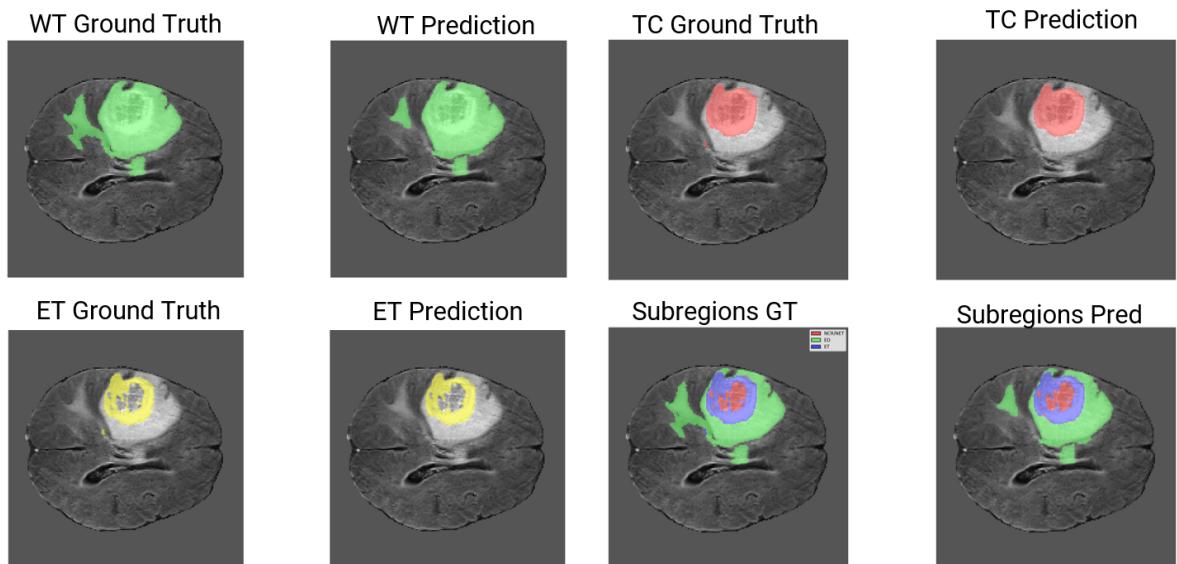
Từ các hình ảnh trên ta thấy mô hình UNETR phân đoạn rất chính xác cả ba vùng WT, TC

và ET trên bệnh nhân 91 và 11, với sự trùng khớp cao giữa prediction và ground truth. Đặc biệt, vùng ET (màu vàng) - thường khó nhất - được xác định tốt ở trung tâm khối u. Kết quả subregions cho thấy mô hình phân biệt rõ ràng ba vùng con NCR/NET (đỏ), ED (xanh lá) và ET (xanh dương).

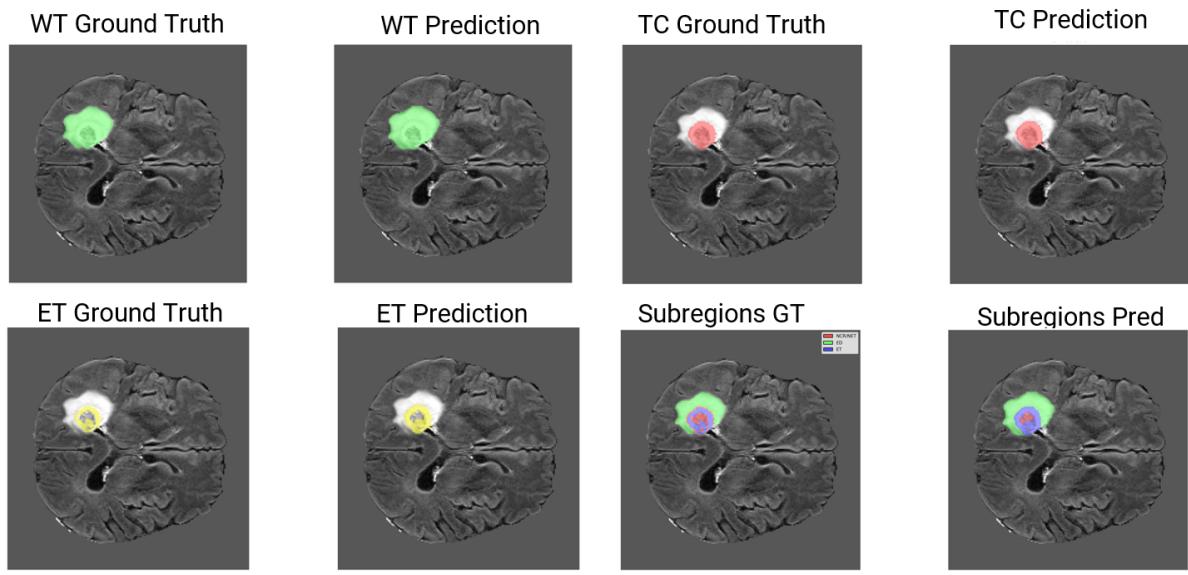
#### □ Mô hình Swin UNet3D

<b>Region</b>	<b>Dice</b>	<b>IoU</b>	<b>ASD</b>	<b>HD95</b>
WT	0.9108	0.8351	1.6035	5.5856
TC	0.8446	0.7639	1.5692	5.3643
ET	0.8084	0.7091	1.1882	4.2647

Bảng 5.5: Kết quả inference mô hình Swin UNet3D trên tập test

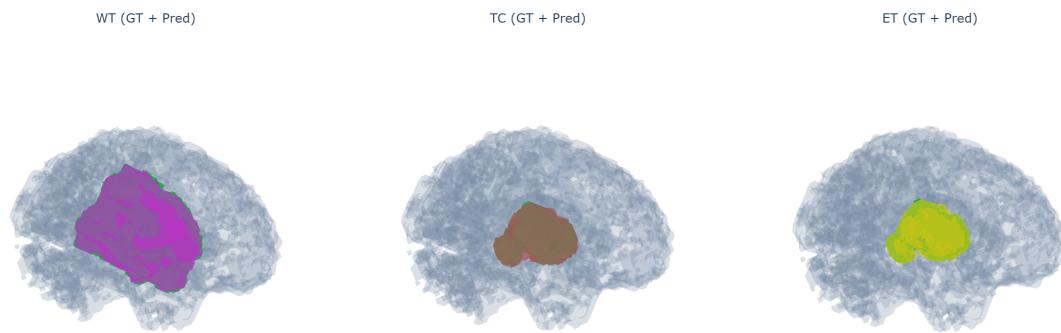


Hình 5.8: Kết quả phân đoạn của Swin UNet3D trên lát cắt thứ 88 của ca chụp 091



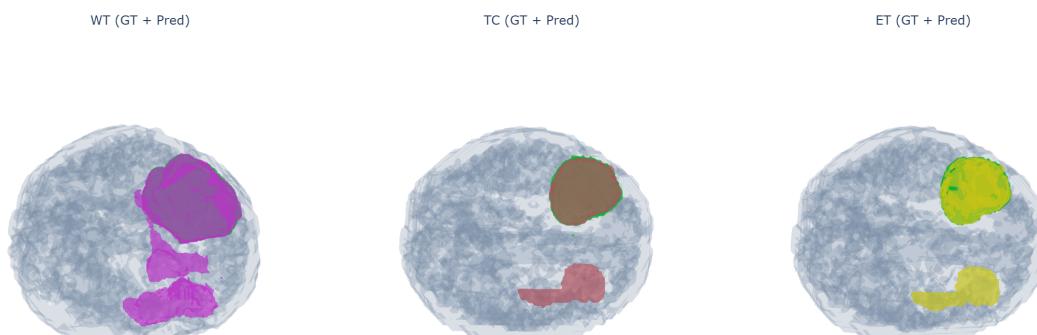
Hình 5.9: Kết quả phân đoạn của Swin UNet3D trên lát cắt thứ 75 của ca chụp 011

Minh họa trường hợp mô hình dự đoán tốt:



Hình 5.10: Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 104

Minh họa trường hợp mô hình dự đoán tệ:



Hình 5.11: Kết quả dựng mesh 3D kết quả dự đoán và ground truth (màu xanh lá) của ca chụp 193

## 5.2 So sánh các phương pháp

Bảng 5.6: So sánh các phương pháp trên ROI WT

<b>Phương pháp</b>	<b>Dice ↑</b>	<b>IoU ↑</b>	<b>ASD ↓</b>	<b>HD95 ↓</b>
UNet	0.8612	0.7725	3.6368	19.0979
UNet++	0.8954	0.8158	<b>1.5725</b>	<b>5.0218</b>
VNet (3D)	0.9024	0.8236	2.0219	6.9727
UNETR	0.8729	0.7887	3.4713	10.5030
Swin UNet3D	<b>0.9108</b>	<b>0.8351</b>	<i>1.6035</i>	<i>5.5856</i>

Bảng 5.7: So sánh các phương pháp trên ROI TC

<b>Phương pháp</b>	<b>Dice ↑</b>	<b>IoU ↑</b>	<b>ASD ↓</b>	<b>HD95 ↓</b>
UNet	0.7549	0.6570	4.6777	12.5280
UNet++	0.8196	0.7297	2.0064	6.2421
VNet (3D)	<b>0.8633</b>	<b>0.7664</b>	1.6688	5.5709
UNETR	0.7819	0.6777	5.2696	12.7303
Swin UNet3D	<b>0.8446</b>	<b>0.7639</b>	<b>1.5692</b>	<b>5.3643</b>

Bảng 5.8: So sánh các phương pháp trên ROI ET

<b>Phương pháp</b>	<b>Dice ↑</b>	<b>IoU ↑</b>	<b>ASD ↓</b>	<b>HD95 ↓</b>
UNet	0.7148	0.5992	9.1472	18.2997
UNet++	0.7360	0.6278	1.6917	5.0079
VNet (3D)	0.7617	0.6650	1.6565	5.3642
UNETR (3D)	0.6869	0.5788	4.2527	12.4837
Swin UNet3D	<b>0.8084</b>	<b>0.7091</b>	<b>1.1882</b>	<b>4.2647</b>

### Nhận xét:

Nhìn chung, các mô hình tiên tiến hơn như UNet++, VNet 3D, UNETR và đặc biệt là Swin UNet3D đều cho kết quả vượt trội so với UNet cơ bản, khẳng định vai trò của việc cải tiến kiến trúc và khai thác ngữ cảnh không gian 3D.

Đối với vùng Whole Tumor (WT), Swin UNet3D đạt hiệu năng tốt nhất với Dice và IoU cao nhất, đồng thời duy trì sai số biên (ASD, HD95) ở mức thấp. Điều này cho thấy khả năng mô hình hóa ngữ cảnh toàn cục và thông tin đa tỉ lệ của kiến trúc Swin Transformer giúp cải thiện đáng kể độ chính xác phân đoạn toàn khối u. VNet huấn luyện trên các patch 3D với Dice Loss cũng cho kết quả cạnh tranh, đặc biệt về Dice và IoU, tuy nhiên kém hơn Swin UNet3D ở độ

chính xác biên.

Đối với vùng Enhancing Tumor (ET), là vùng khó phân đoạn nhất do kích thước nhỏ và độ tương phản không ổn định, Swin UNet3D vượt trội trên tất cả các chỉ số đánh giá. Kết quả này cho thấy mô hình có khả năng học được các đặc trưng ngữ cảnh dài hạn và mối quan hệ không gian phức tạp, giúp cải thiện đáng kể chất lượng phân đoạn đối với các vùng u nhỏ.

So với UNet truyền thống, UNet++ cho thấy sự cải thiện nhất quán trên cả ba vùng ROI, đặc biệt ở các chỉ số đo độ chính xác biên (HD95), chứng minh hiệu quả của các kết nối dense skip trong việc thu hẹp khoảng cách ngữ nghĩa giữa encoder và decoder. UNETR, mặc dù khai thác Transformer thuần túy, vẫn chưa đạt hiệu năng cao bằng các mô hình lai CNN–Transformer trong bối cảnh dữ liệu hạn chế.

Tổng thể, kết quả thực nghiệm cho thấy Swin UNet3D là mô hình cho hiệu năng toàn diện và ổn định nhất trên hầu hết các vùng ROI và chỉ số đánh giá. Điều này khẳng định tiềm năng của các kiến trúc kết hợp Transformer và CNN trong bài toán phân đoạn khối u não 3D, đặc biệt khi yêu cầu độ chính xác cao ở ranh giới và các vùng u nhỏ.