

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R. Roth, and Daguang Xu

2022.12.20 논문 리뷰

배성훈

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Research Background:**

- 최근의 Medical Image segmentation에서 FCNN기반 방법론은 좋은 성능을 보임
- 하지만 FCNN기반 방법론은 다음과 같은 문제점을 가짐
 1. CNN-based 방법은 **제한된 kernel size** 때문에 **Long-range dependencies**의 학습에 제한이 있고, 다양한 size와 shape으로 나타내어지는 tumors의 정확한 segmentation을 **제한**한다.
 2. 대안으로 사용된 **Receptive field 확장**은 **Local region**에 대한 **한계**를 보임
- 저자는 이러한 한계를 **Long-range dependencies**와 **Multi-scale features**를 잘 추출하고 학습하는 UNETR에 영감을 받아, swin transformer를 결합해 새로운 **Swin UNETR**이라는 모델을 제안.

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Research Background:**

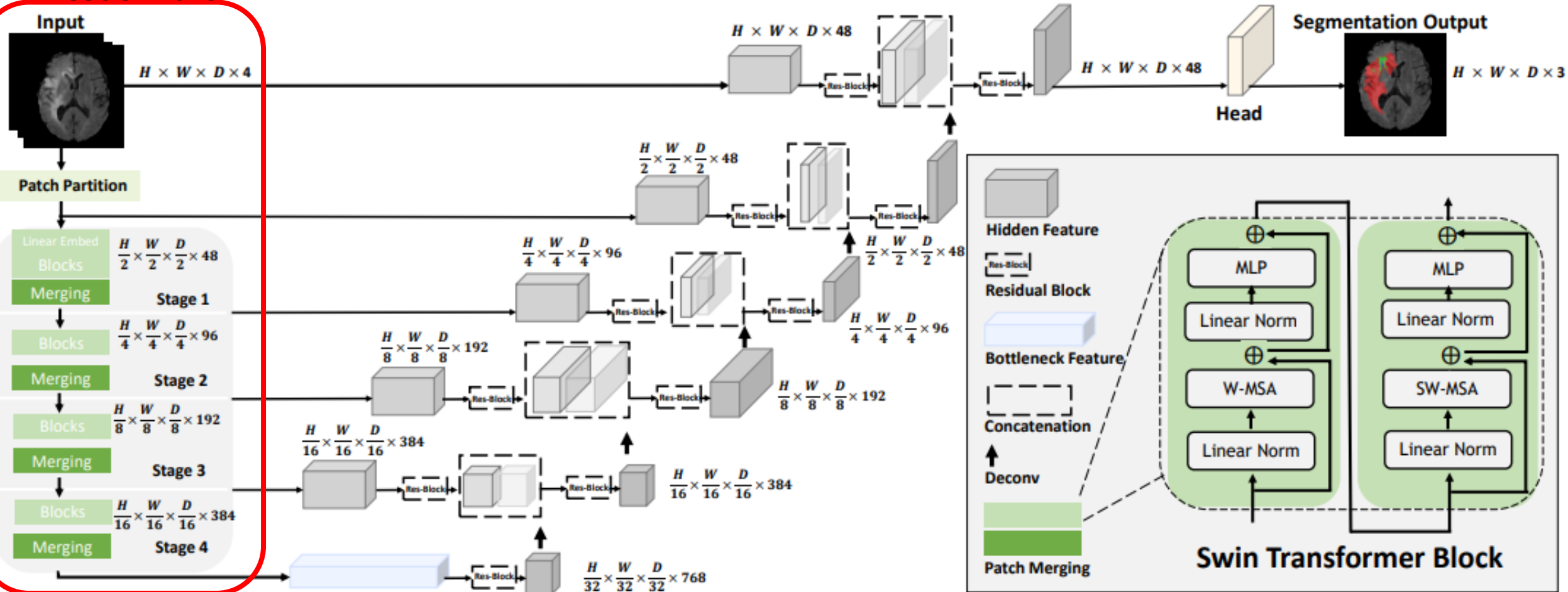
- Swin UNETR은 Voxel 각각에 대해 굉장히 민감한 Task인 medical segmentation에서 기존의 UNETR을 개선한 방식
- UNETR에서 사용한 ViT 방식의 경우 고해상도 Image에 대해서 **Quadratic하게 증가하는 연산량**으로 인해서 고해상도 이미지를 그대로 사용할 수 없거나, 학습에 **오랜 시간 및 많은 비용**이 드는 단점을 가짐
- 이러한 문제를 **shifted window** 방법을 통해 **linear하게 연산량을 증가**시켜 해결함
- 또한 resolution 복원을 위해 CNN-based decoder를 통해 서로 다른 크기의 features를 skip connection을 통해 영상 확장시 추가 정보로 활용
- 결과적으로, **Multi-scale contextual representation**의 학습과 **long-range dependencies**를 가능하게 함

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

• Architecture: Encoder

- Patch partition, Linear Embedding, Swin Transformer Block, Patch Merging,
- Swin Transformer Block은 2개의 Encoder로 구성되어있고, Window Multi-Head self-attention, Shifted Window Multi-Head Self-Attention으로 구성

Encoder Part

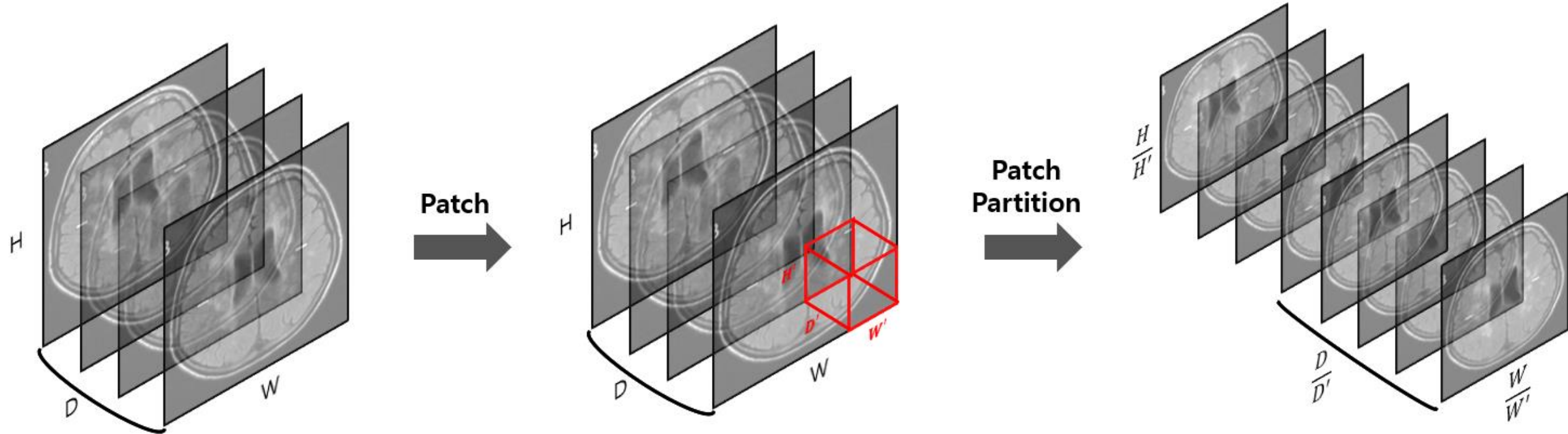


Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Patch Partition

- Input = $X \in \mathbb{R}^{H \times W \times D \times S}$ 을 (H', W', D') 의 patch 크기로 token화 시킴 (Patch size = $2 \times 2 \times 2$)
- $X \in \mathbb{R}^{\frac{H}{H'} \times \frac{W}{W'} \times \frac{D}{D'} \times C}$

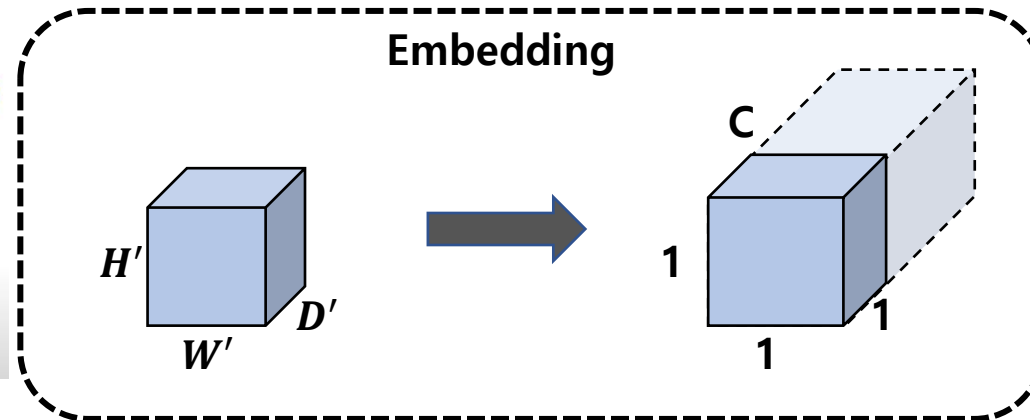
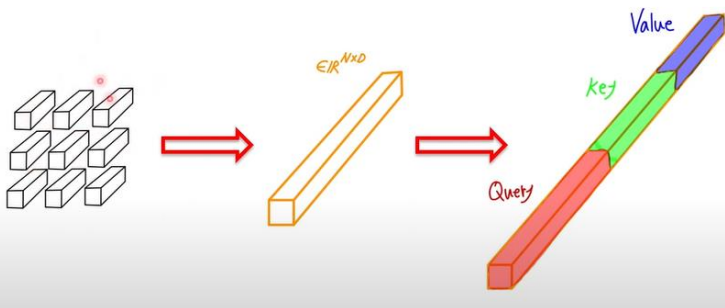
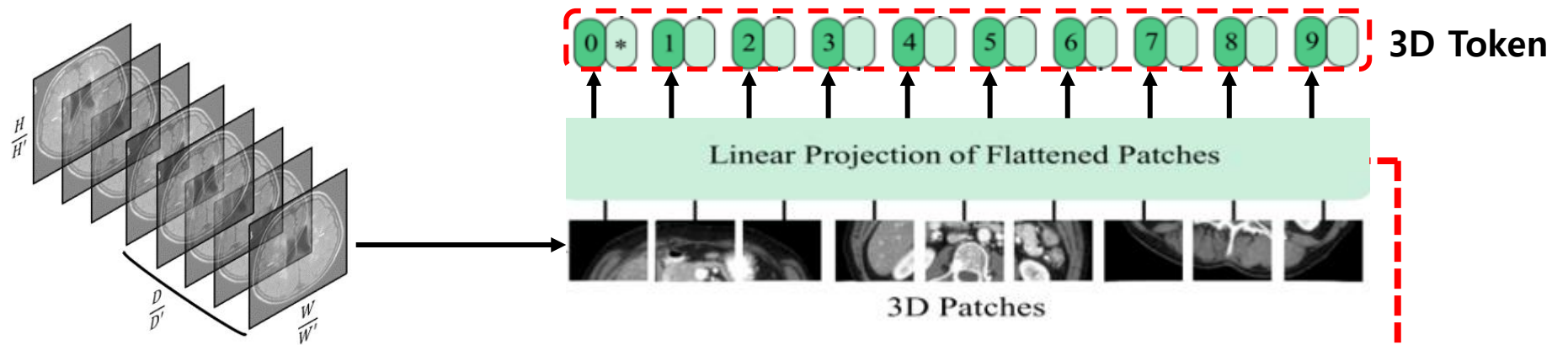
Patch Partition



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

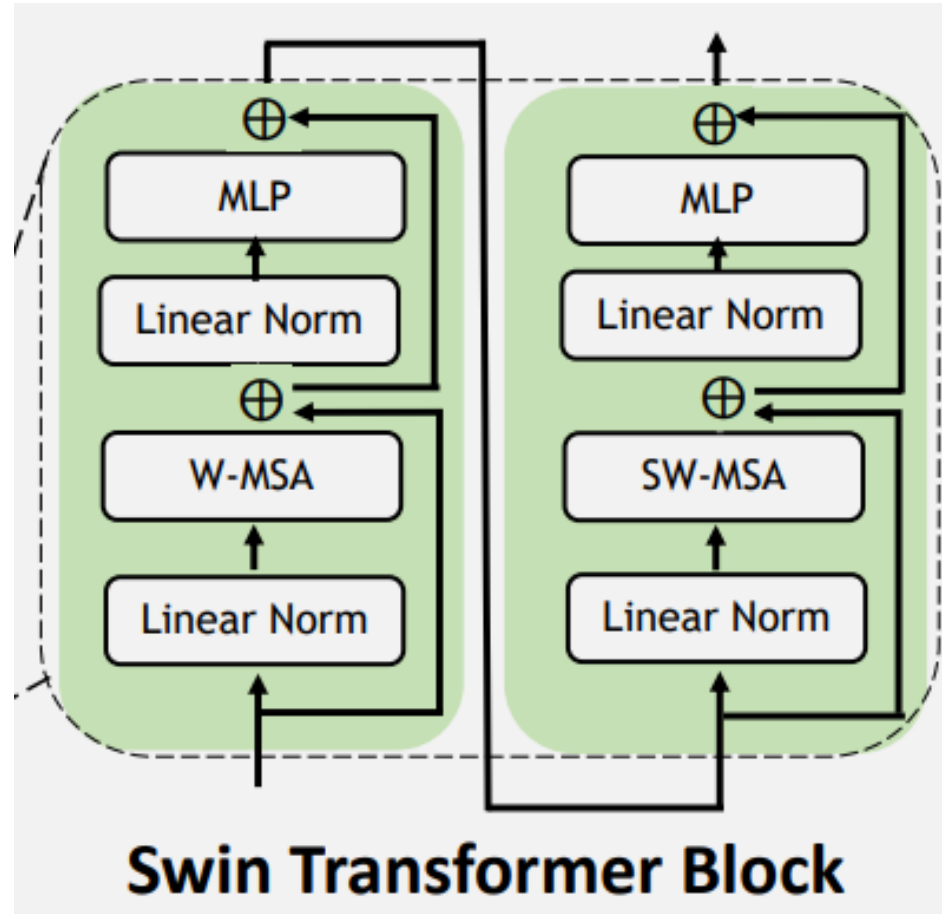
- **Method:** Linear Embedding
 - Patch Partition의 출력의 각 **Patch**들은 Linear Embedding layer를 거쳐 C의 dimension으로 embedding **C=(48, 96, 192, 384)**
 - Stage마다 Embedding space가 2배씩 증가

Linear Embedding



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

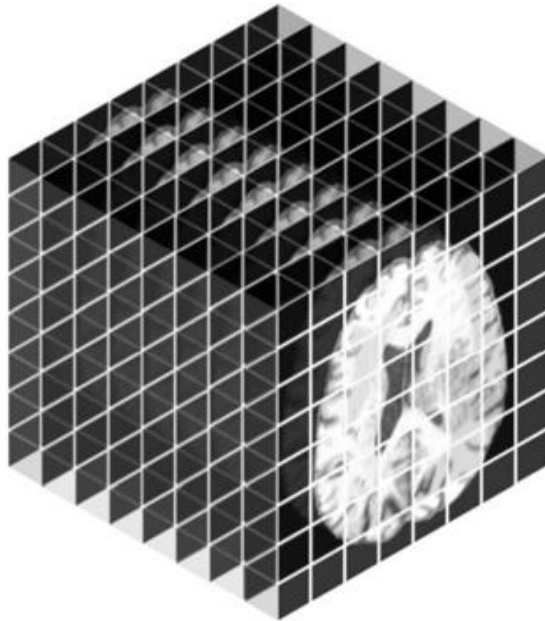
- **Method:** Swin Transformer Block
 - Linear embedding의 출력을 입력으로 사용
 - $layer\ l, layer\ l + 1$ 로 나뉘서 학습



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

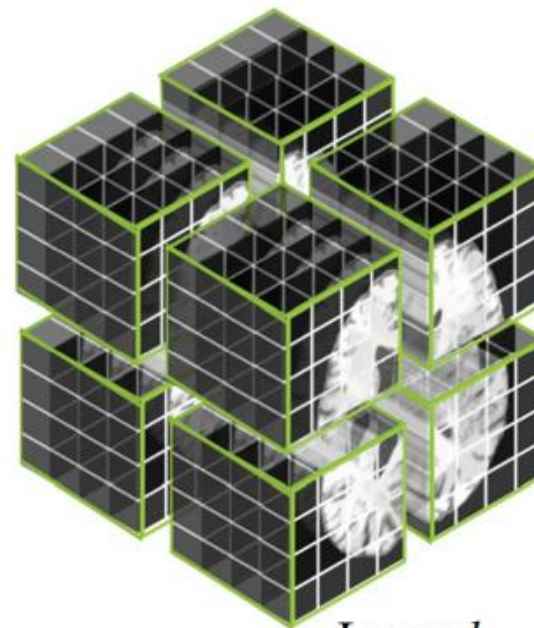
- **Method:** Swin Transformer Block(*layer l*)

- Linear embedding 출력을 사용
- $M \times M \times M$ 크기의 window size로, 입력된 linear embedding을 $\left\lceil \frac{H'}{M} \right\rceil \times \left\lceil \frac{W'}{M} \right\rceil \times \left\lceil \frac{D'}{M} \right\rceil$ 영역으로 파티션.



3D Tokens: $8 \times 8 \times 8$

Window size: $4 \times 4 \times 4$



Layer *l*

Number of windows: 8

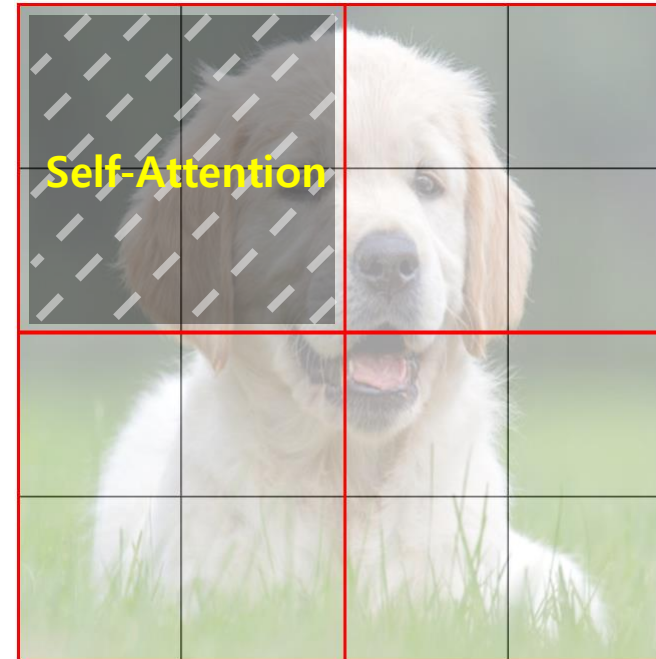
Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block(*layer l*)

- **Window Multihead Self-attention (W-MSA)**

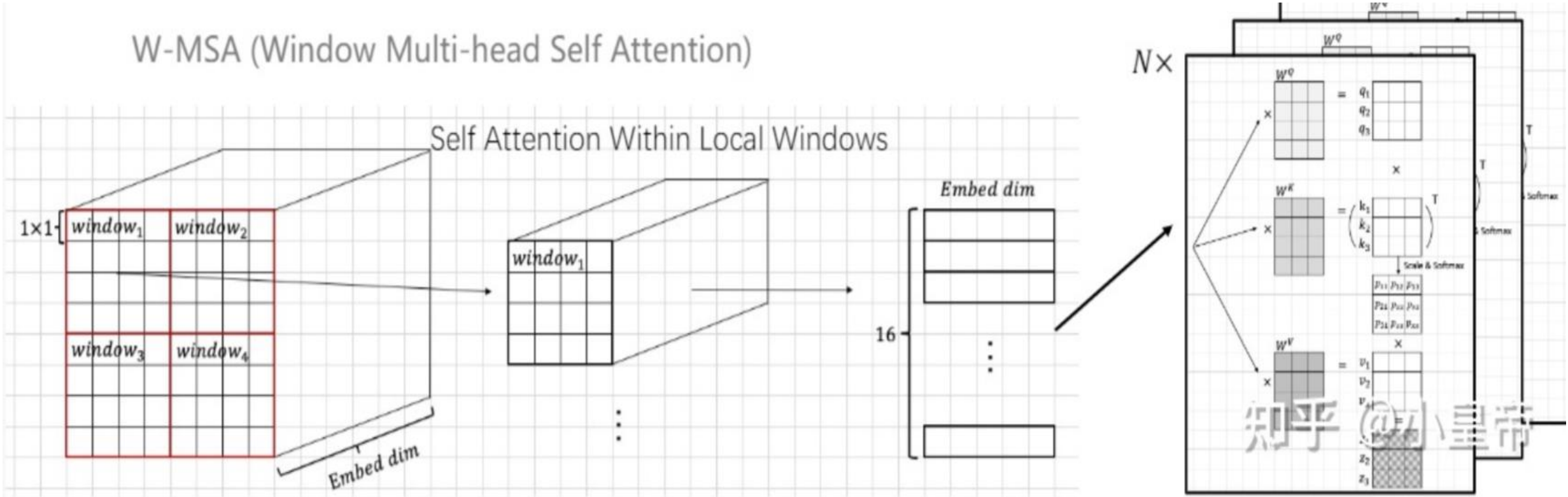
- 현재 window에 있는 패치끼리 만 Self-Attention 수행
 - 이는 주변 픽셀들끼리 서로 연관성을 높여 computational complexity를 해결 (**Quadratic -> Linear**)
 - **M(Window size) < hw(image size)** 때문에 **W-MSA의 연산량 < MSA의 연산량**

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C,$$
$$\Omega(\text{W-MSA}) = 4hwC^2 + 2\underline{M^2}hwC,$$



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block(*layer l*)
Window Multihead Self-attention (W-MSA)

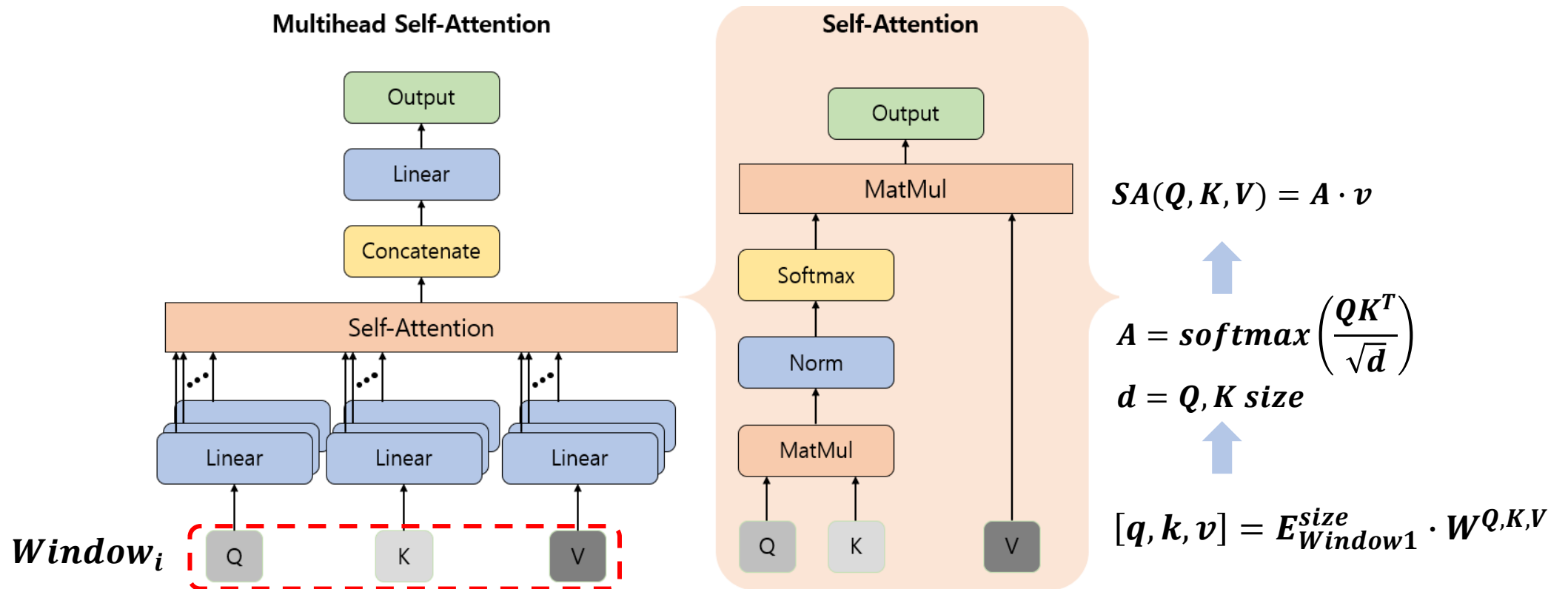


Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block(*layer l*)

- **Window Multihead Self-attention (W-MSA)**

- Window 각각에 대한 Multihead Self-attention 적용
 - $SA(Q, K, V)$ 유닛을 N개 concat



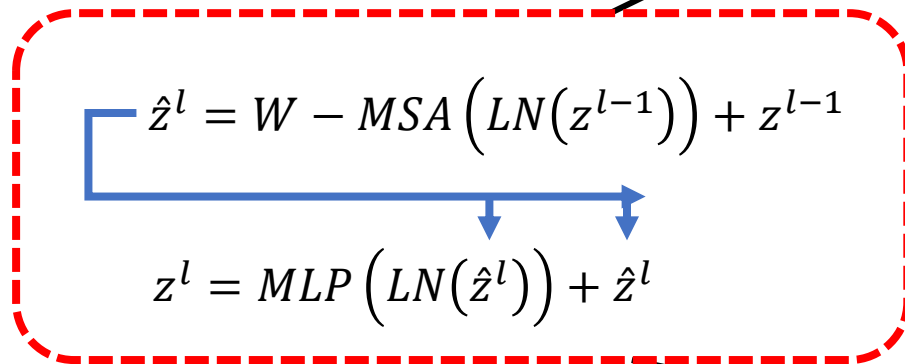
Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block(*layer l*)

- W-MSA를 거친 feature representation은 최초의 입력된 linear embedding 출력과 concat한 후 **MLP** 과정을 거침

- **Method:** Swin Transformer Block(*layer l*)

- W-MSA를 거친 feature representation은 최초의 입력된 linear embedding 출력과 concat한 후 **MLP** 과정을 거침

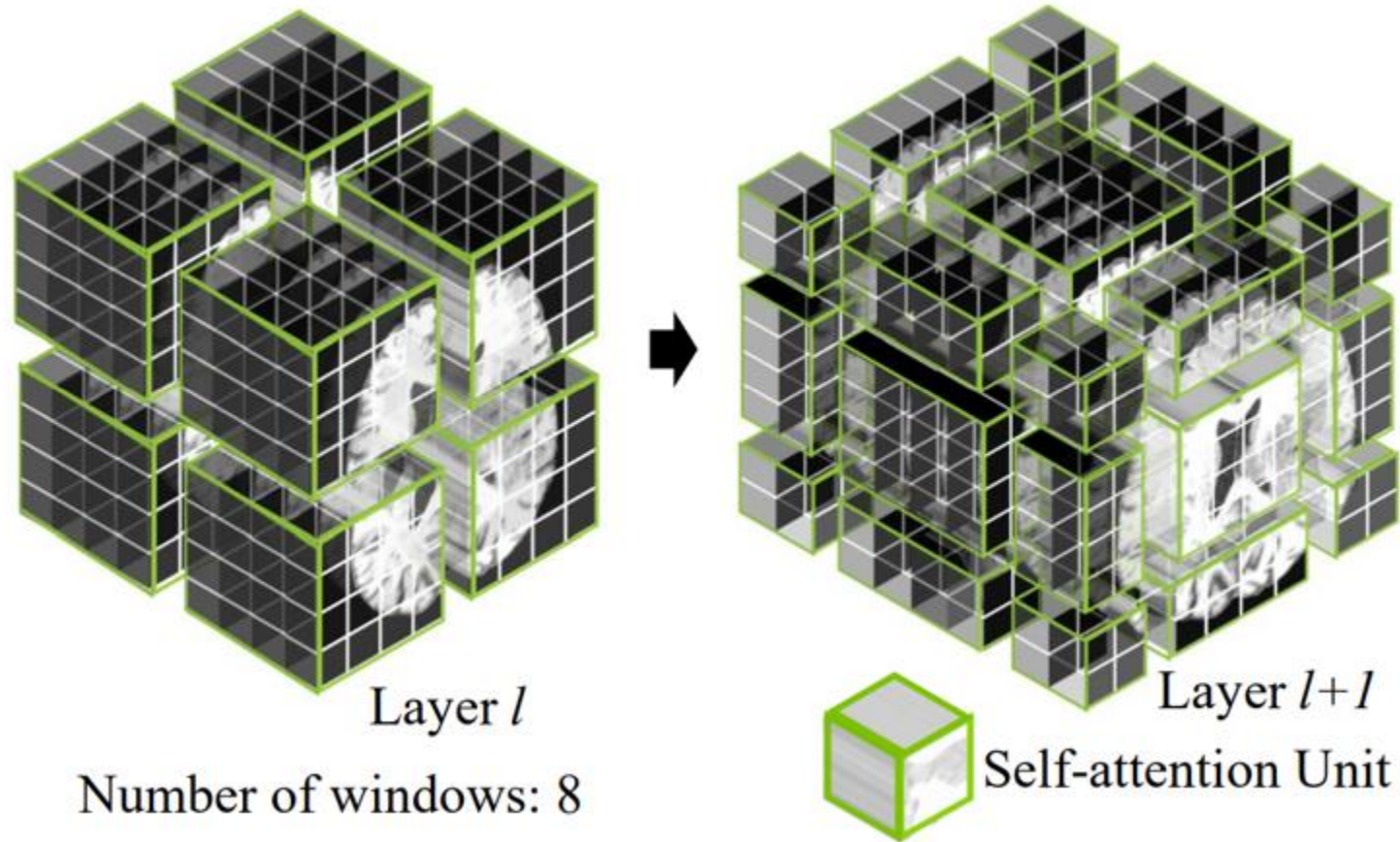


The diagram illustrates the internal structure of a Swin Transformer Block for layer l . It is enclosed in a red dashed rounded rectangle. A blue line starts from the left, goes up, then right, and then down to point at the first equation. The first equation is $\hat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1}$. From the right side of this equation, a blue line goes right and then down to point at the second equation. The second equation is $z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l$. Two black arrows originate from the text blocks above: one points to the top of the red dashed box, and the other points to the bottom of the red dashed box.

$$\hat{z}^l = W - MSA(LN(z^{l-1})) + z^{l-1}$$
$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l$$

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block($layer\ l + 1$)
 - $layer\ l$ 의 출력을 입력으로 사용
 - $\lceil \frac{M}{2} \rceil, \lceil \frac{M}{2} \rceil, \lceil \frac{M}{2} \rceil$ voxels 크기만큼 window regions를 shift함



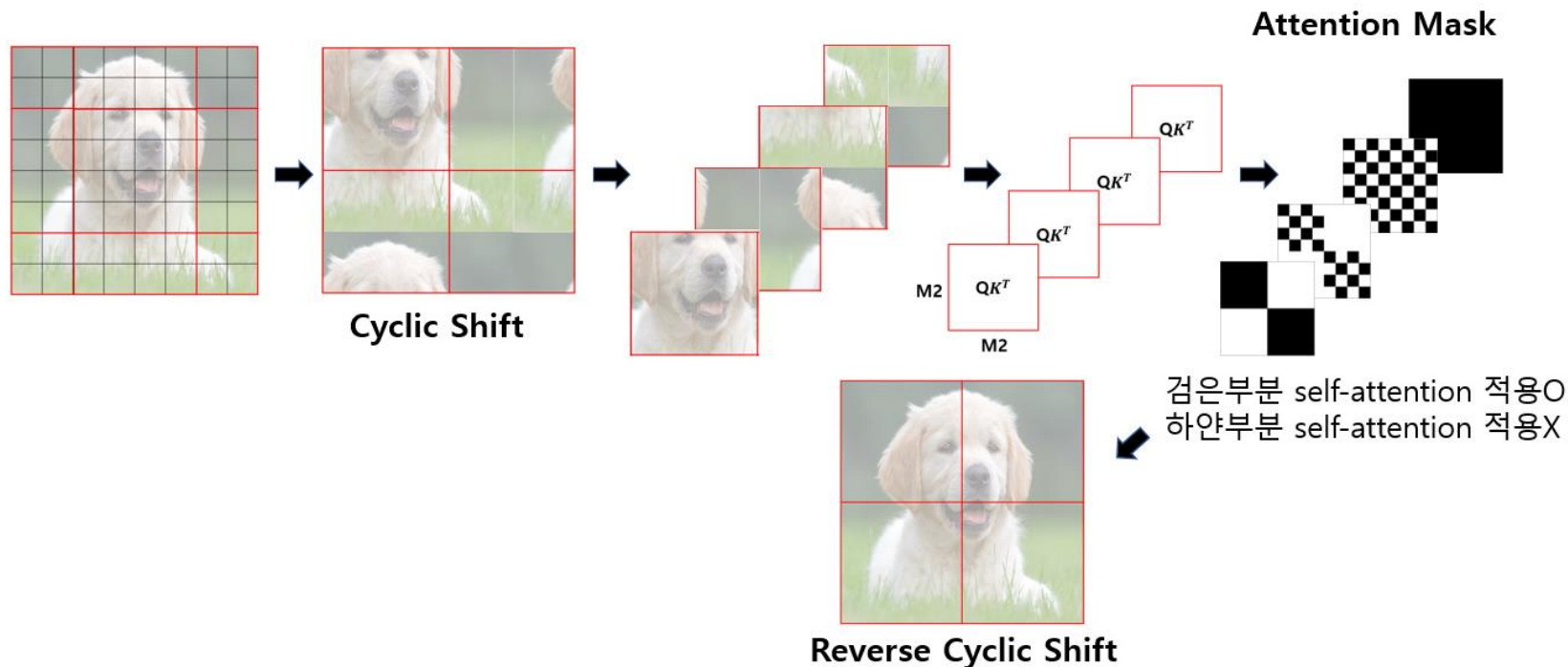
Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block(*layer l*)

Shifted Window Multihead Self-attention (SW-MSA)

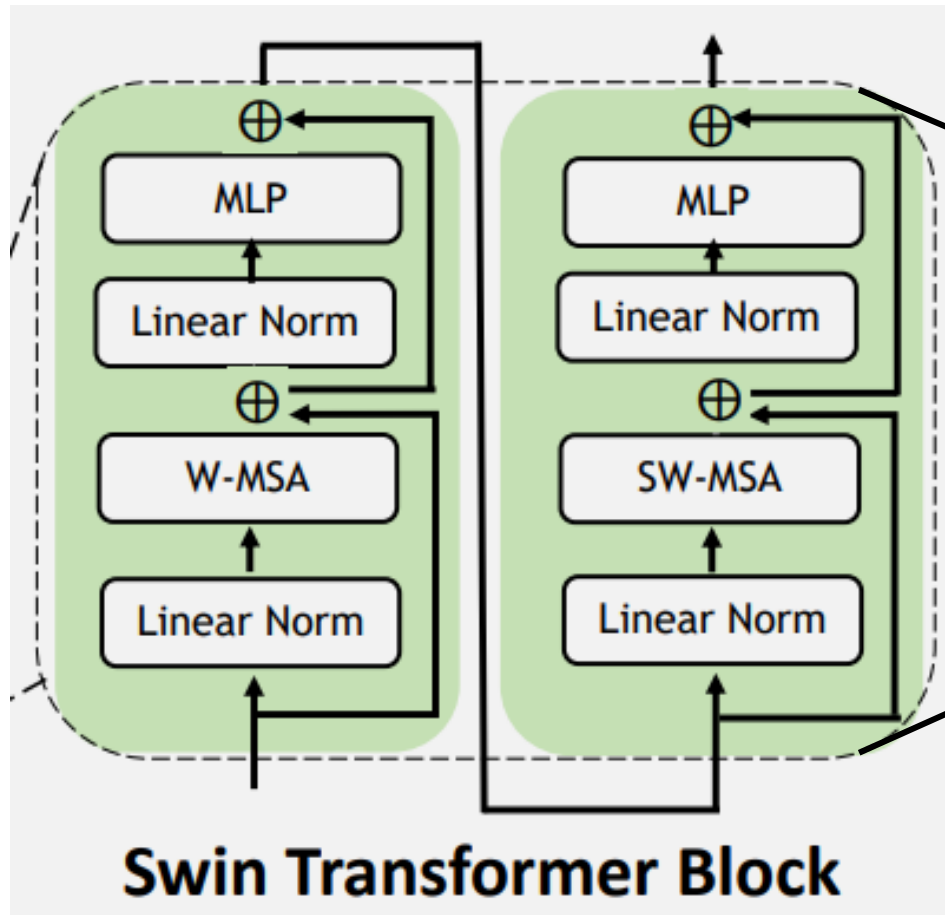
- 하지만 W-MSA는 Window를 고정해, **고정된 부분**에서만 Self-attention을 수행해 **멀리 떨어진 Patch들과 상호작용이 불가능**
- 이를 해결하기 위해, Window간의 연결을 추가하면서 적은 연산량을 유지하는 **Shifted Window Multihead Self-attention** 제안

Cyclic Shift와 Masked MSA 예시



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Swin Transformer Block(*layer l + 1*)
 - SW-MSA를 거친 feature representation은 *layer l*의 출력과 concat한 후 **MLP** 과정을 거침

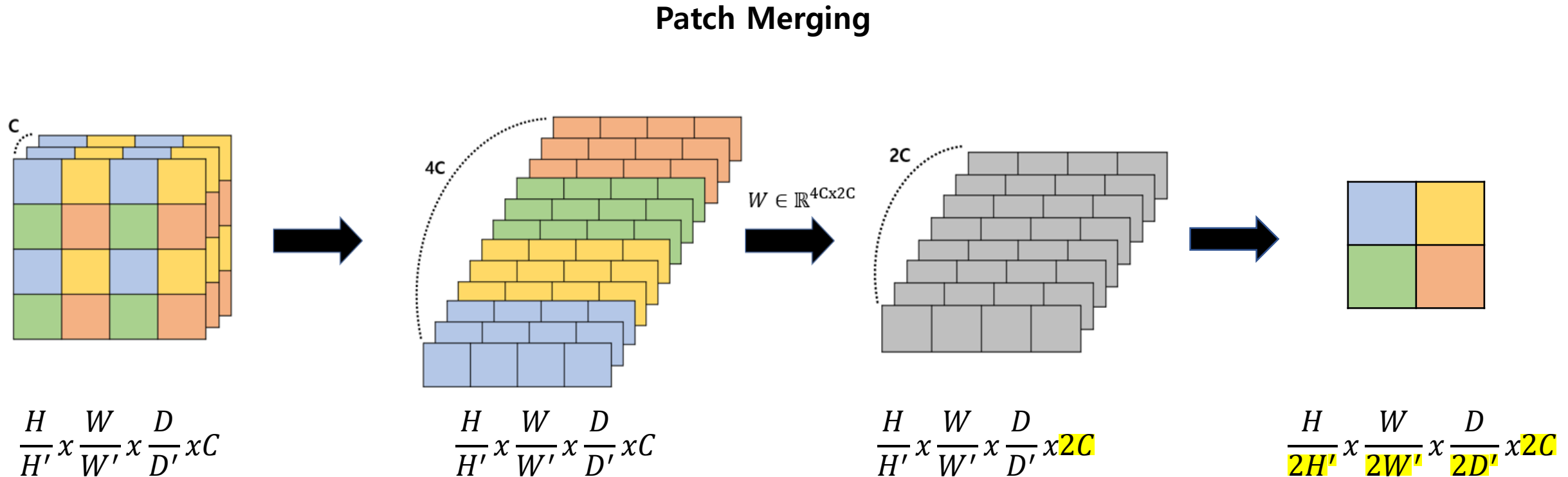


$$\hat{z}^{l+1} = SW - MSA \left(LN(z^l) \right) + z^l$$
$$z^{l+1} = MLP \left(LN(\hat{z}^{l+1}) \right) + \hat{z}^{l+1}$$

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Patch Merging Layer

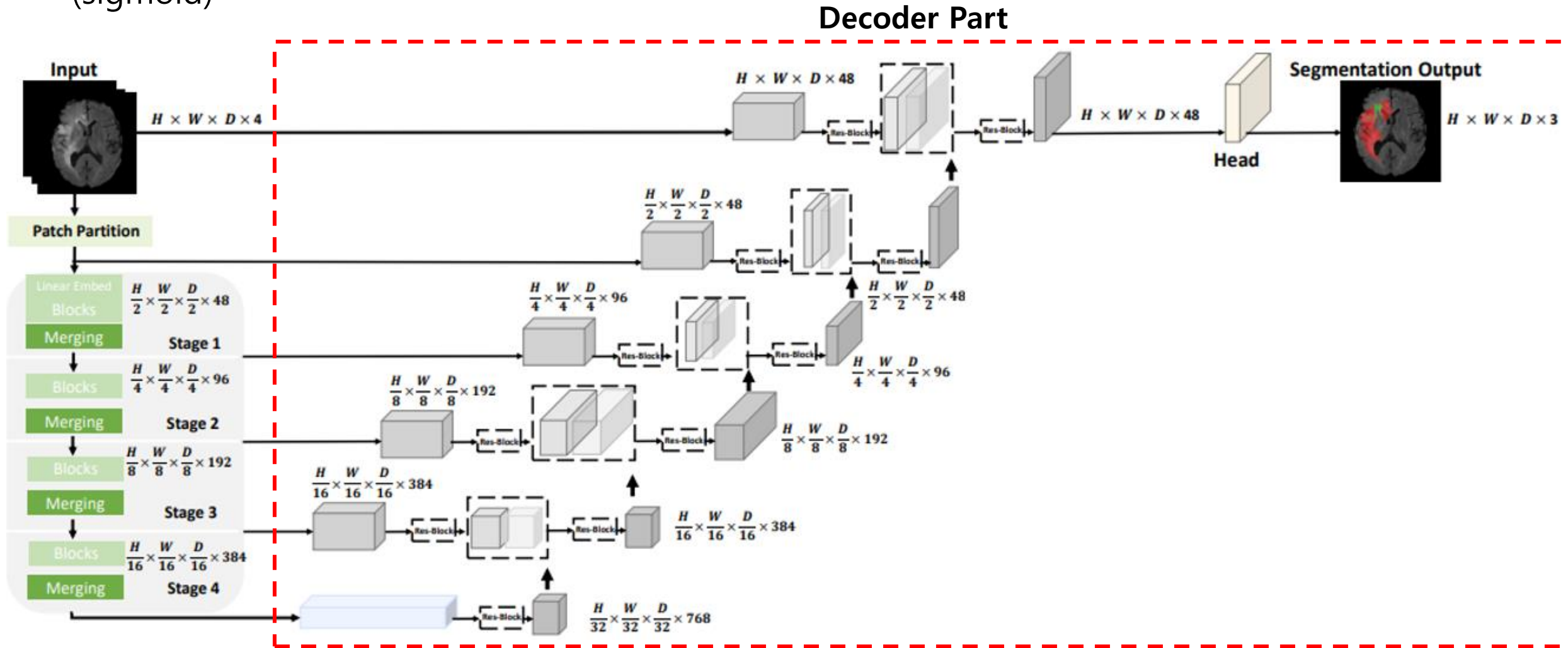
- Model의 hierarchical 구조를 위해 patch merging에서 각 stage의 출력을, factor=2로 feature representation 감소!!
- + 2x2x2 크기의 patches를 그룹화하고 concat 함



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Architecture:** Decoder

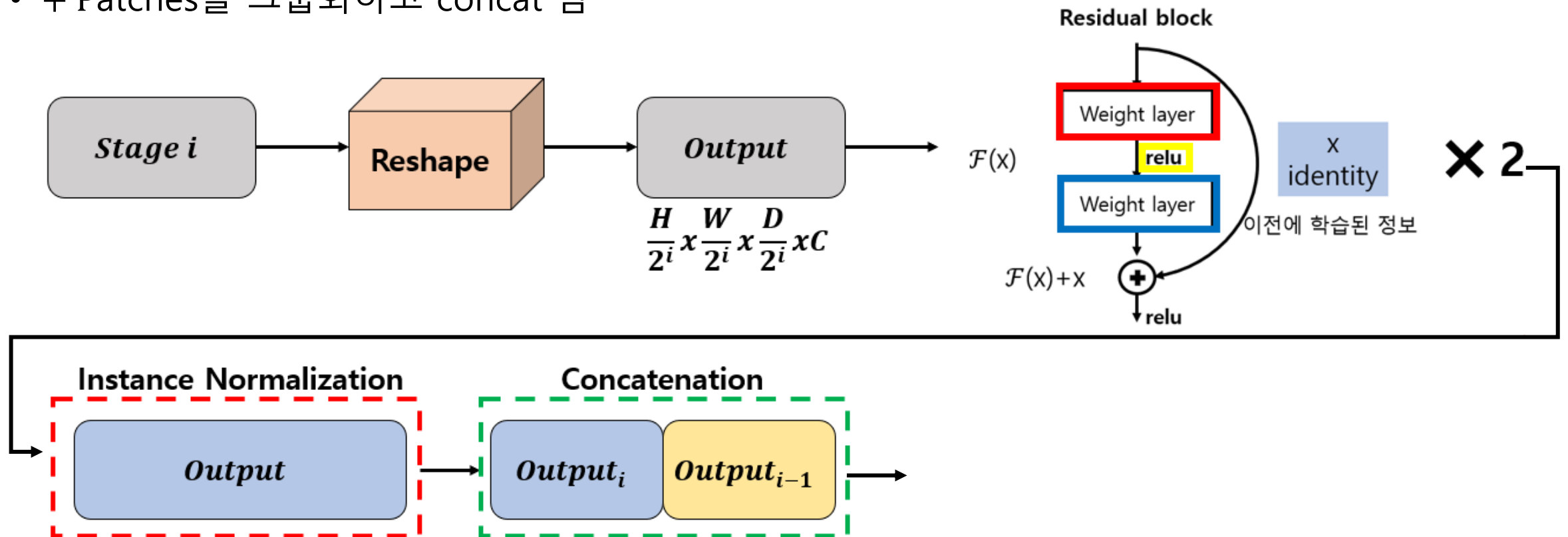
- Reshape($\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}$), 3x3x3 Residual block x2 (instance normalization), Concatenation
- 3x3x3 Residual block x2 (instance normalization), Deconvolution(factor=2), 1x1x1 convolution (sigmoid)



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Skip Connection & Concatenation

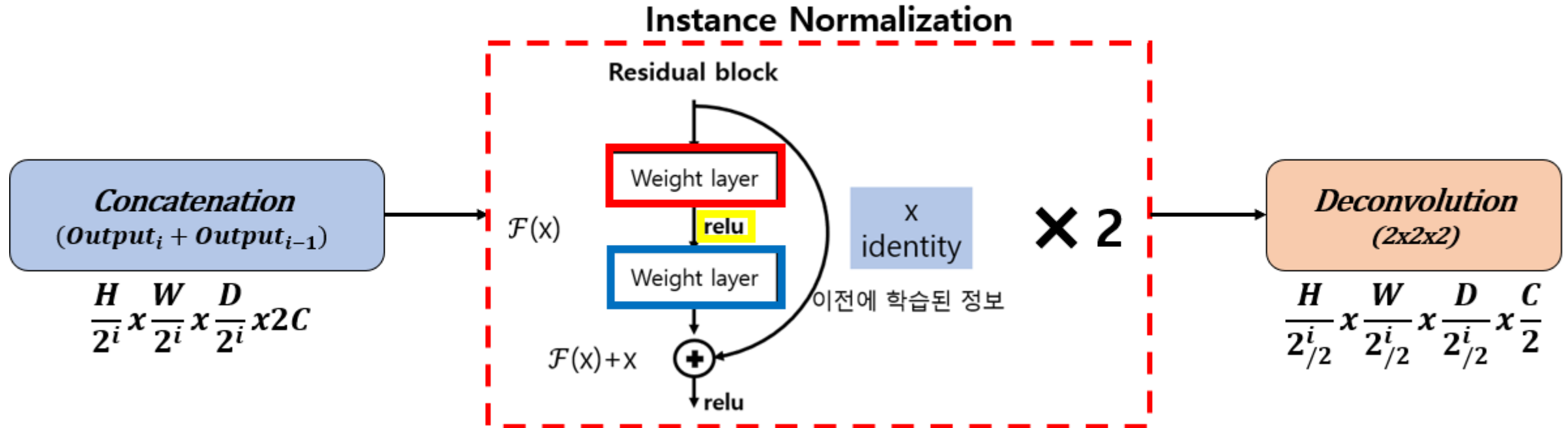
- 1. 각 stage i ($i \in \{0,1,2,3,4,5\}$, $5 = \text{bottleneck}$) 의 output feature representation O_i $\frac{H}{2^i} \times \frac{W}{2^i} \times \frac{D}{2^i}$ 의 크기로 reshape됨
- 2. Reshape된 output feature representation은 2개의 $3 \times 3 \times 3$ Residual block과 함께 Instance normalization을 거침
- 3. 2번 과정이 완료된 output은 이전 stage의 output과 concat 과정을 거침
- + Patches를 그룹화하고 concat 함



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

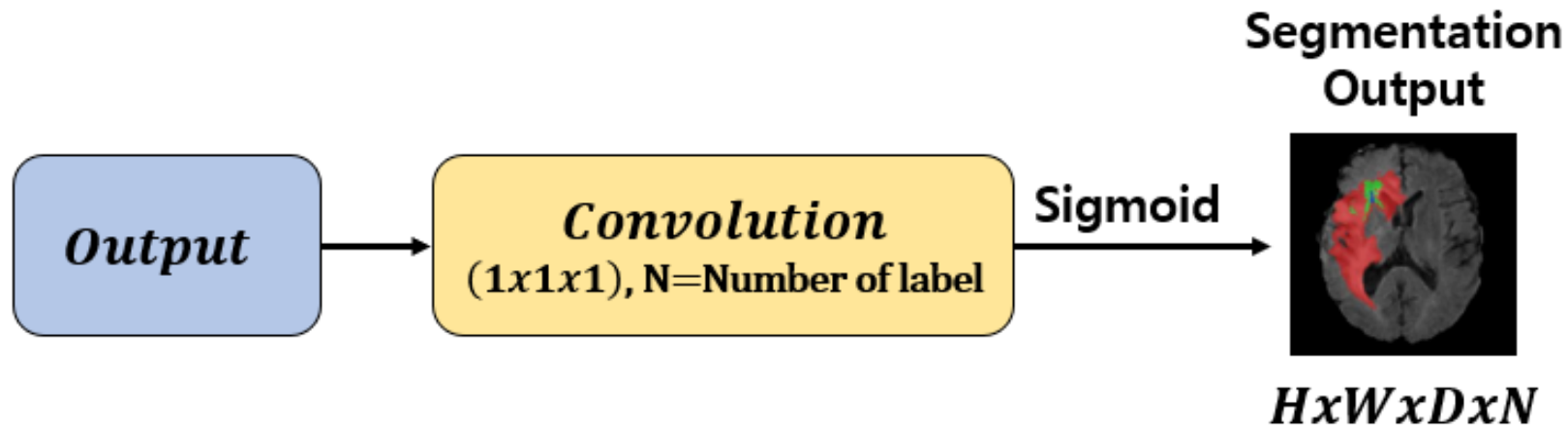
- **Method:** Deconvolution

- 1. Concat된 output은 2개의 3x3x3 residual block을 거친 후 Instance normalization으로 정규화 과정을 진행
- 2. Feature representation은 2x2x2 Deconvolution을 통해 2배 확장



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:** Segmentation Head
 - 전체 과정을 거친 후 feature representation은 $1 \times 1 \times 1$ convolution layer를 거친 후 **Sigmoid activation function**을 통해 최종 Segmentation output을 계산
 - N = number of label



Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Method:**

- **Loss function**

- Soft Dice loss function

$$L(G, Y) = 1 - \frac{2}{J} \sum_{j=1}^J \frac{\sum_{i=1}^I G_{i,j} Y_{i,j}}{\sum_{i=1}^I G_{i,j}^2 + \sum_{i=1}^I Y_{i,j}^2}$$

- I : voxels numbers
 - J : Classes number
 - $G_{i,j}$: Probability of output
 - $Y_{i,j}$: Probaility of one-hot encoded ground truth

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- **Implementation:**
 - **Learning rate** : 0.0008
 - **Data Normalization** : zero-mean, unit standard deviation
 - zero-mean, unit standard deviation: $x_i - \left(\frac{x_{mean}}{x_{std}}\right)$
 - 데이터 정규화를 통해 데이터 전체 샘플 범위를 일정하게 조정하고 분산을 1로 고정
 - **Data Augmentation**
 - **Cropped** 128x128x128 Random patches
 - Random axis mirror **flip** (all 3 axes)
 - Random **channel intensity shift**(-0.1, 0.1), **scale intensity**(0.9, 1.1))
 - **Epoch** : 800
 - **Cosine annealing learning rate scheduler**
 - 학습율의 최대값과 최소값을 정해서 그 범위의 학습율을 코사인 함수를 이용하여 스케줄링하는 방법
 - 모델 일반화 극대화
 - **Sliding window** with overlapping of 0.7
 - **Five fold cross-validation** (80:20)

Swin UNETR: Swin Transformer for Semantic Segmentation of Brain Tumors in MRI Images

- Experiments:
 - Five fold cross-validation 모든 결과에서 기존의 방식보다 더 좋은 성능을 달성
 - 3개의 semantic classes (ET, WT, TC)와 평균에서 기존보다 **0.7%, 0.6%, 0.4%, 0.5%** 더 성능 향상
 - Self attention**과 **long-range dependencies**를 효율적으로 모델링 한 Hierarchical Encoder를 통해 **Multi-scale contextual information**의 학습을 가능하게 함

Five-fold cross-validation benchmarks in terms of mean Dice score values

	Swin UNETR				nnU-Net				SegResNet				TransBTS			
Dice Score	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.	ET	WT	TC	Avg.
Fold 1	0.876	0.929	0.914	0.906	0.866	0.921	0.902	0.896	0.867	0.924	0.907	0.899	0.856	0.910	0.897	0.883
Fold 2	0.908	0.938	0.919	0.921	0.899	0.933	0.919	0.917	0.900	0.933	0.915	0.916	0.885	0.919	0.903	0.902
Fold 3	0.891	0.931	0.919	0.913	0.886	0.929	0.914	0.910	0.884	0.927	0.917	0.909	0.866	0.903	0.898	0.889
Fold 4	0.890	0.937	0.920	0.915	0.886	0.927	0.914	0.909	0.888	0.921	0.916	0.908	0.868	0.910	0.901	0.893
Fold 5	0.891	0.934	0.917	0.914	0.880	0.929	0.917	0.909	0.878	0.930	0.912	0.906	0.867	0.915	0.893	0.892
Avg.	0.891	0.933	0.917	0.913	0.883	0.927	0.913	0.908	0.883	0.927	0.913	0.907	0.868	0.911	0.898	0.891

BraTS 2021 validation dataset benchmarks

	Dice			Hausdorff (mm)		
Validation dataset	ET	WT	TC	ET	WT	TC
Swin UNETR	0.858	0.926	0.885	6.016	5.831	3.770

BraTS 2021 testing dataset benchmarks

	Dice			Hausdorff (mm)		
Testing dataset	ET	WT	TC	ET	WT	TC
Swin UNETR	0.853	0.927	0.876	16.326	4.739	15.309