Part 4
SOTA model

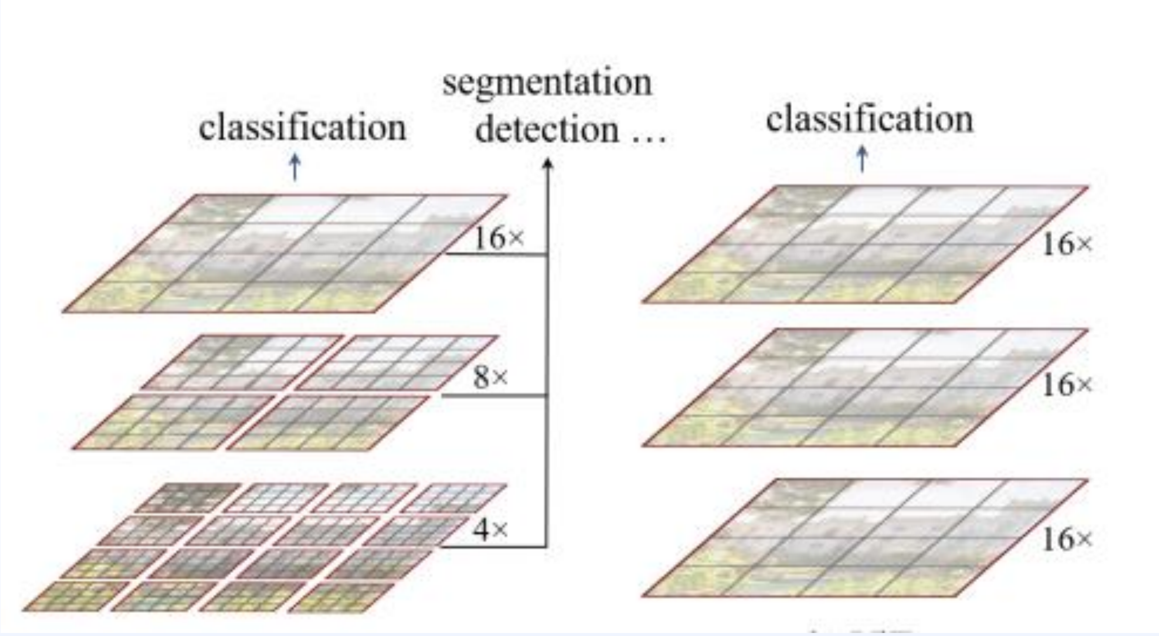
Ch4. Swin Transformer V2

Swin Transformer V2: Scaling Up Capacity and Resolution

Swin Transformer vs ViT

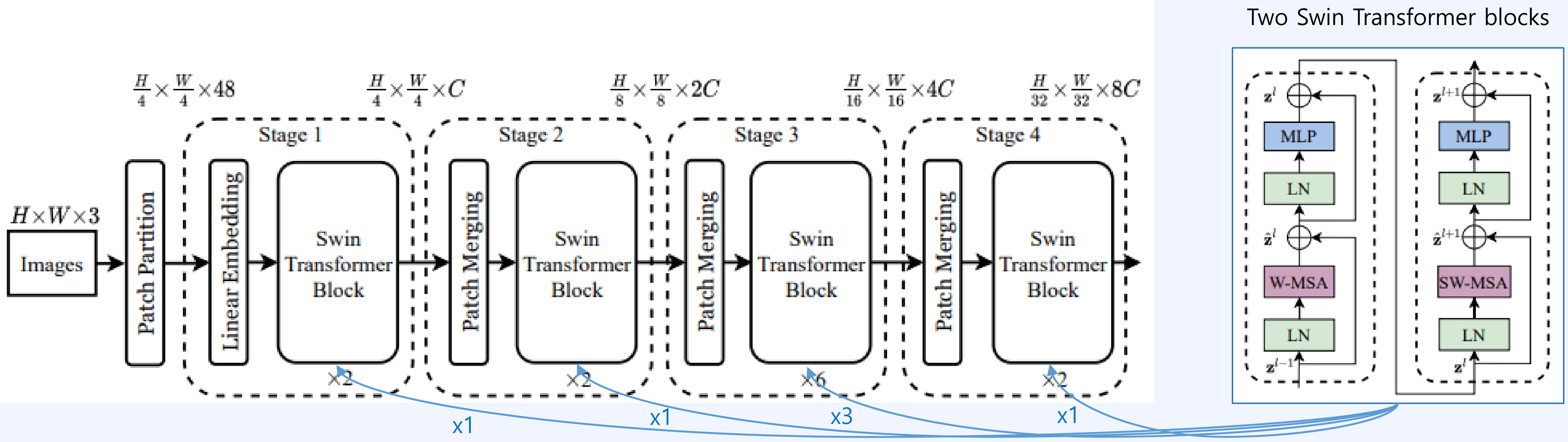
- Swin Transformer
  - 연산량이 window 수에 선형적으로 증가
    - High resolution task 수행
  - Hierarchical representation을 학습
    - Object Detection task 수행
    - Backbone으로 사용

- ViT
  - 연산량이 image 크기의 제곱에 비례
    → High resolution task를 수행 X
  - Hierarchical한 구조X
    → Object Detection task 수행의 어려움
    → backbone의 역할의 어려움



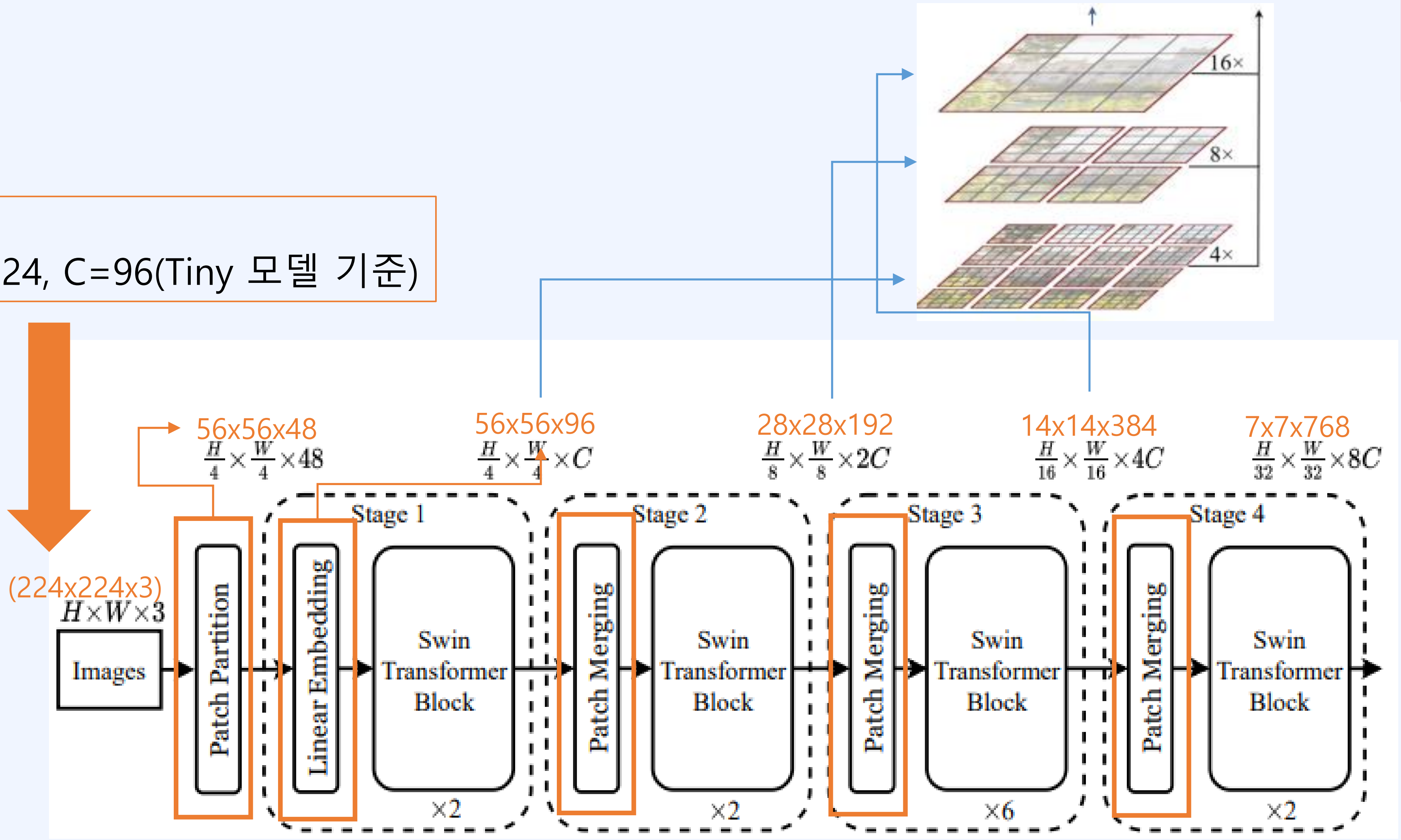
Swin Transformer architecture

Architecture

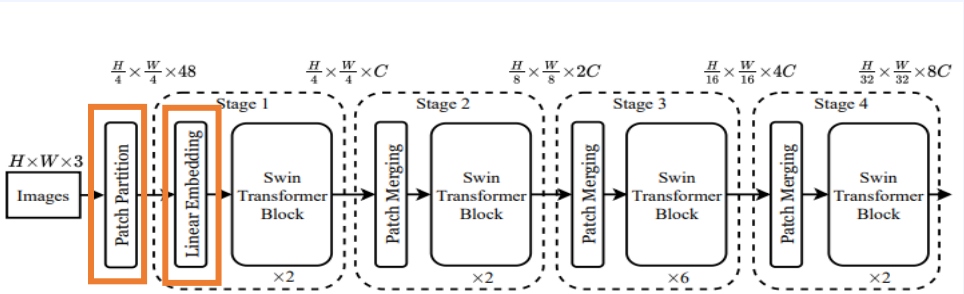


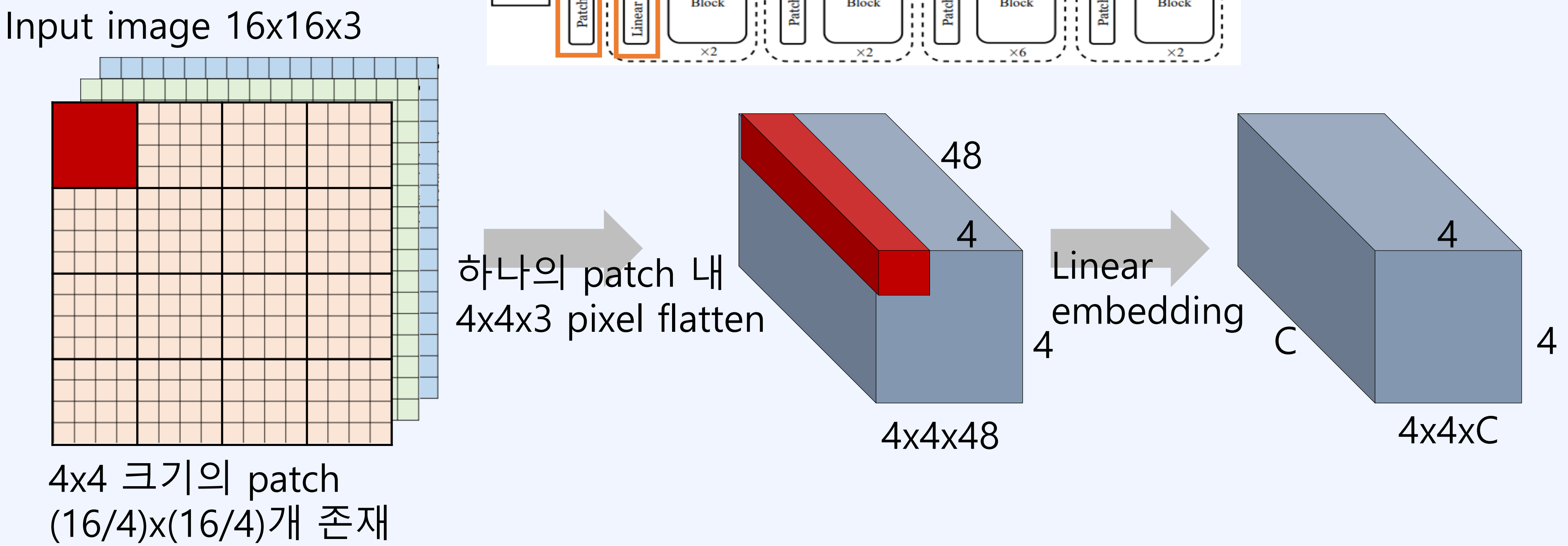
Swin Transformer architecture

<예시>
h=224,w=224, C=96(Tiny 모델 기준)



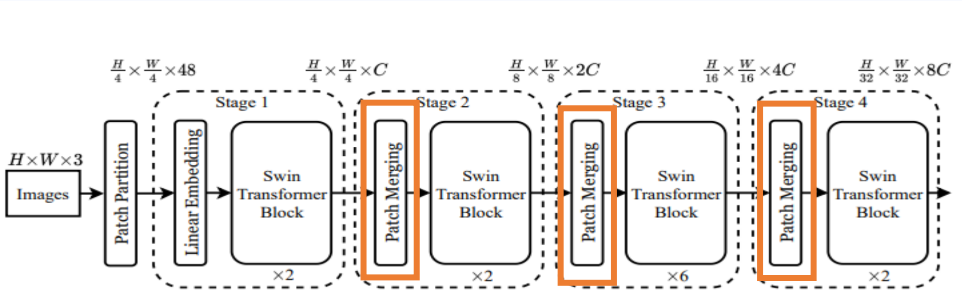
Patch Partition&Linear Embedding

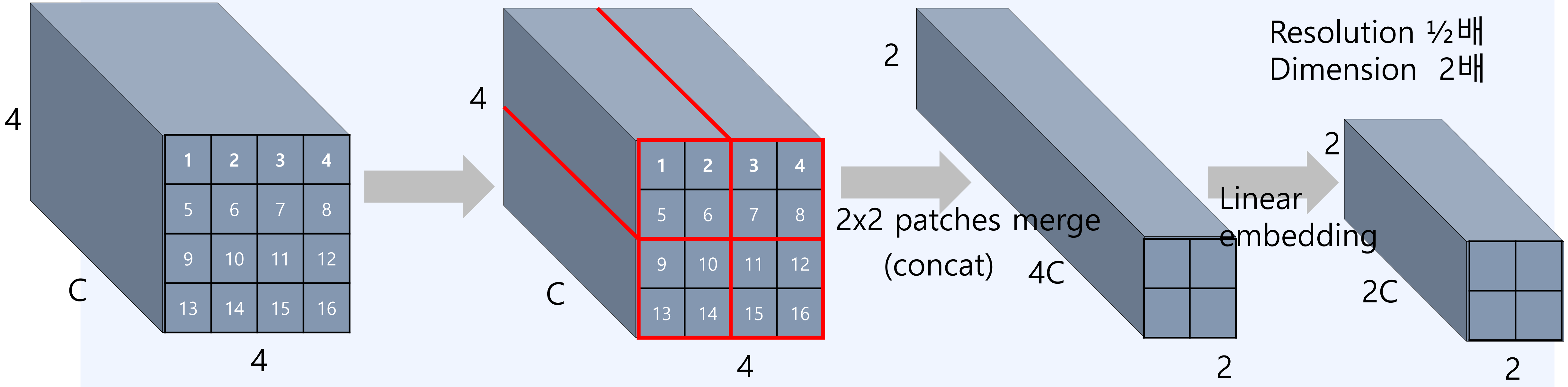




- Patch가 하나의 token
- Patch가 4x4일때, 16x16x3 input image→48차원 1D 16개로 flatten

Patch Merging

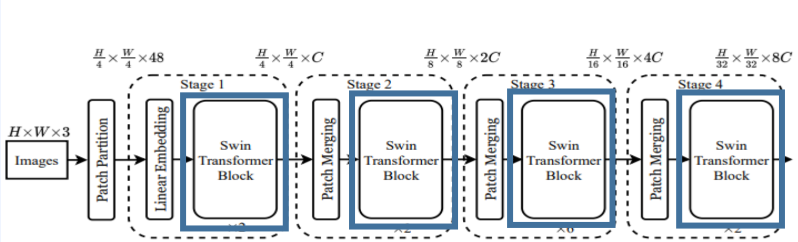


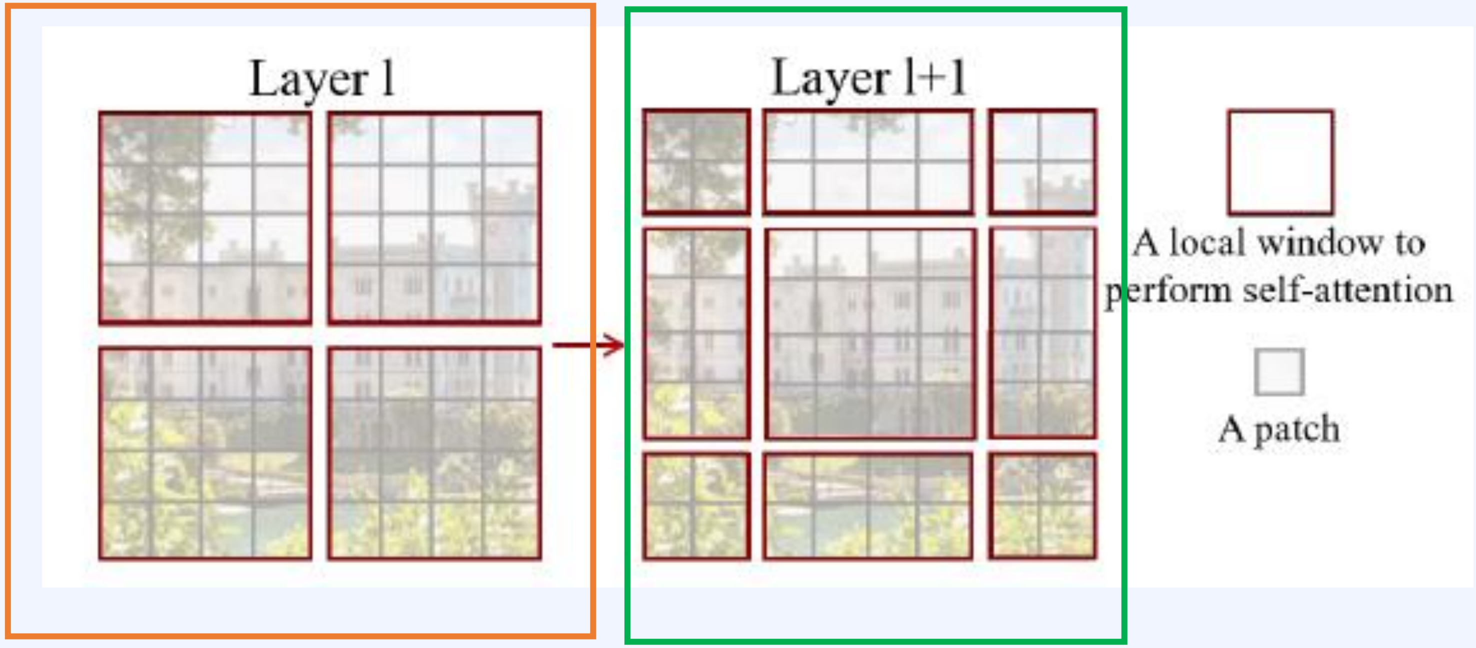


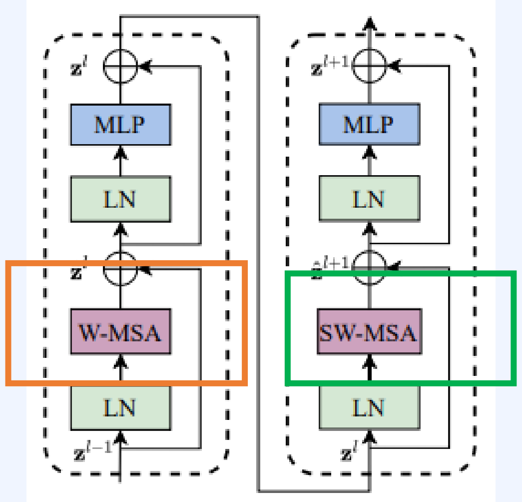
- 2x2 의 neighboring (attention window 내) patch들을 하나의 patch로 concat
- CNN에서 Feature map size를 2배 줄이면, channel 수를 2배로 늘리 것과 비슷

Swin Transformer block

- W-MSA: Local Window 내에서 self attention
- SW-MSA : Local Window 간의 연결성 부여



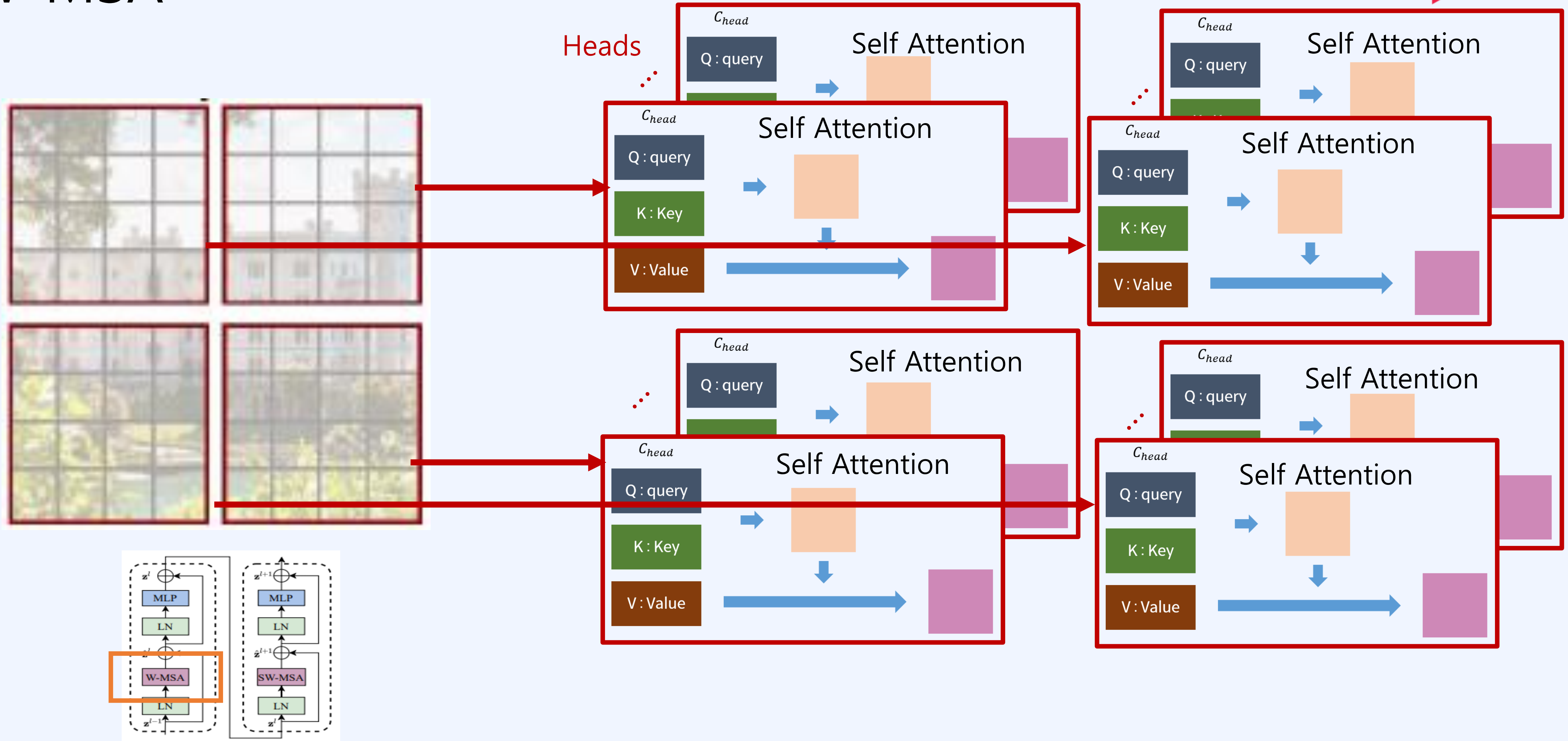




Two Swin Transformer blocks

Swin Transformer block

- W-MSA



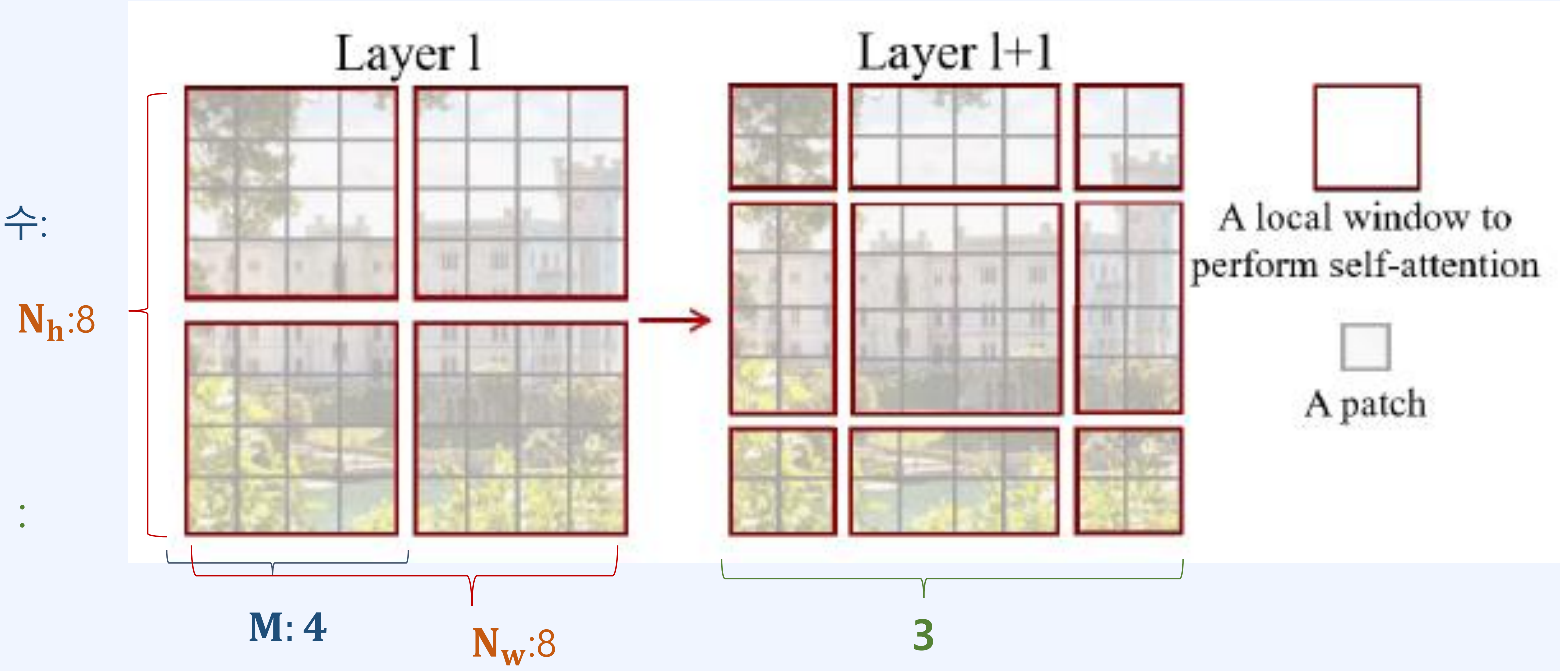
Swin Transformer block

- SW-MSA
  - SW-MSA 실행 시, window 개수: $\left(\frac{N_h}{M}+1\right) \times \left(\frac{N_w}{M}+1\right)$ → window 증가
  - Cyclic Shift 와 Attention Mask를 통해 W-MSA와 동일한 window 개수 사용

Image 하나의 Patch 개수:
$N_h \times N_W$

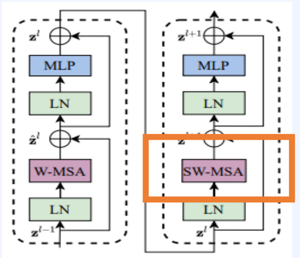
하나의 Winodw내의 patch 개수:
M x M (M=4)

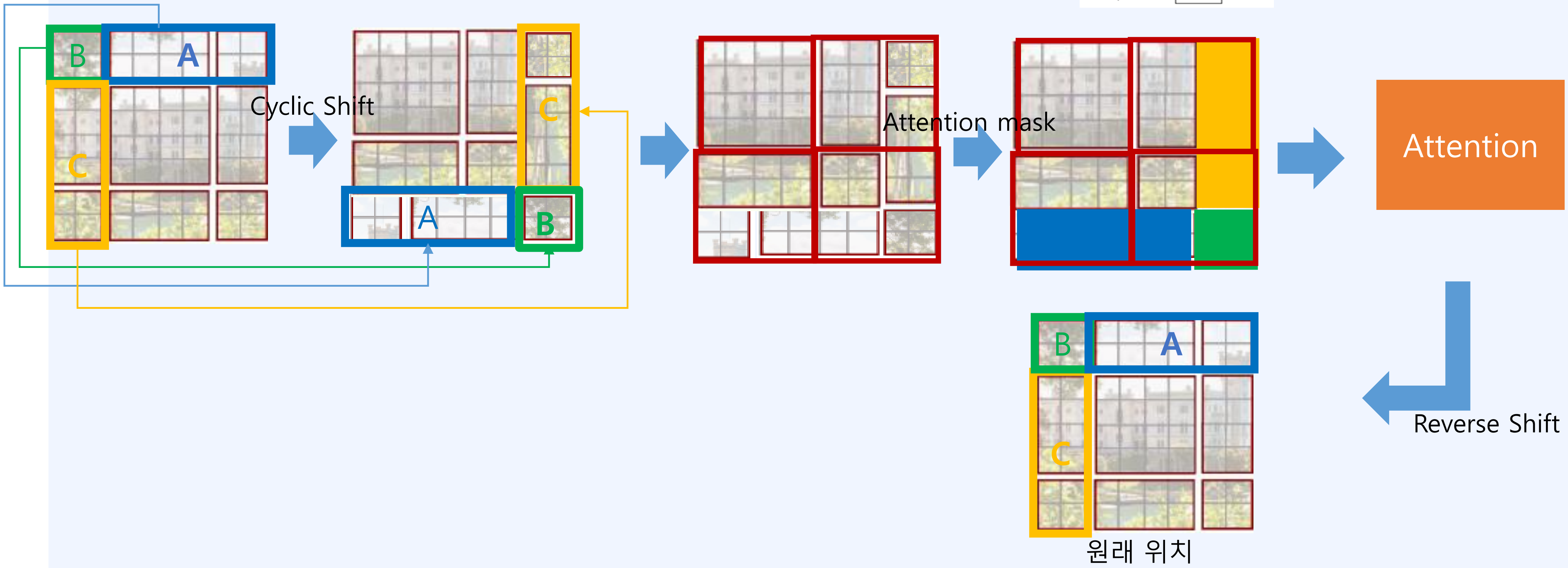
SW-MSA winodow 개수 :
3 x 3



Swin Transformer block

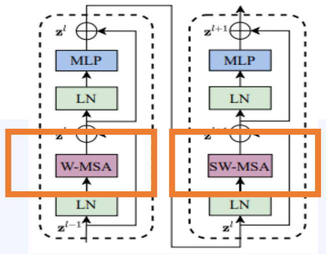
- Cyclic Shift 와 Attention Mask(SW-MSA)





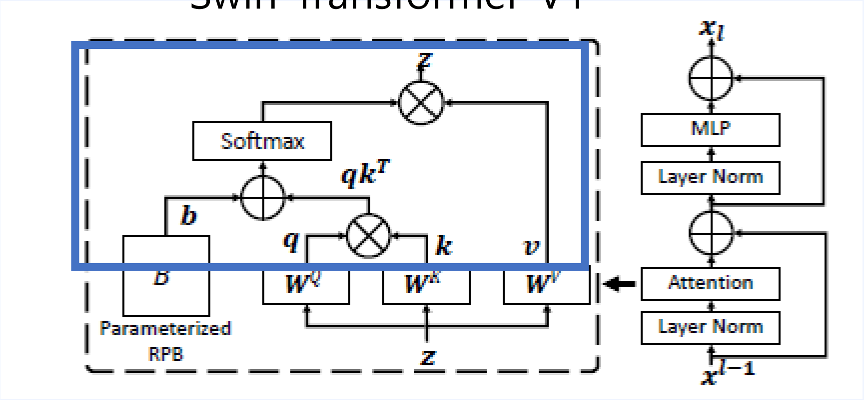
Swin Transformer v2

Swin Transformer block



- Self attention

Swin Transformer V1



C_{head}

Q : query

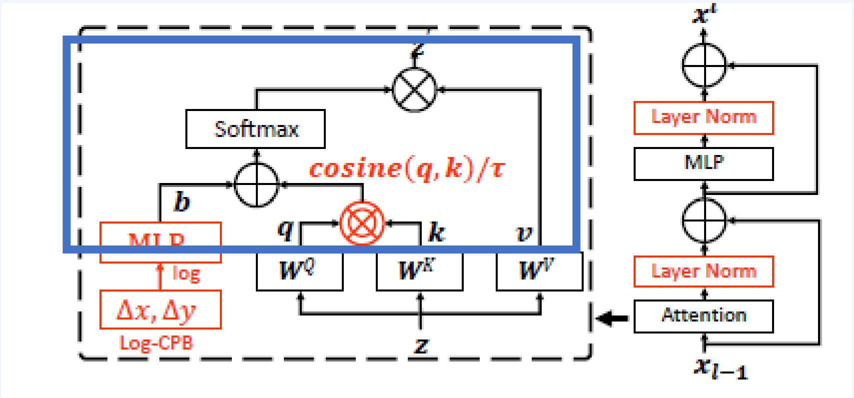
K : Key

V : Value

QK^T

$$softmax\left(\frac{QK^T}{\sqrt{C_{head}}}+b\right)V$$

Swin Transformer V2



C_{head}

Q : query

K : Key

V : Value

$cosine(Q,K)/\tau$

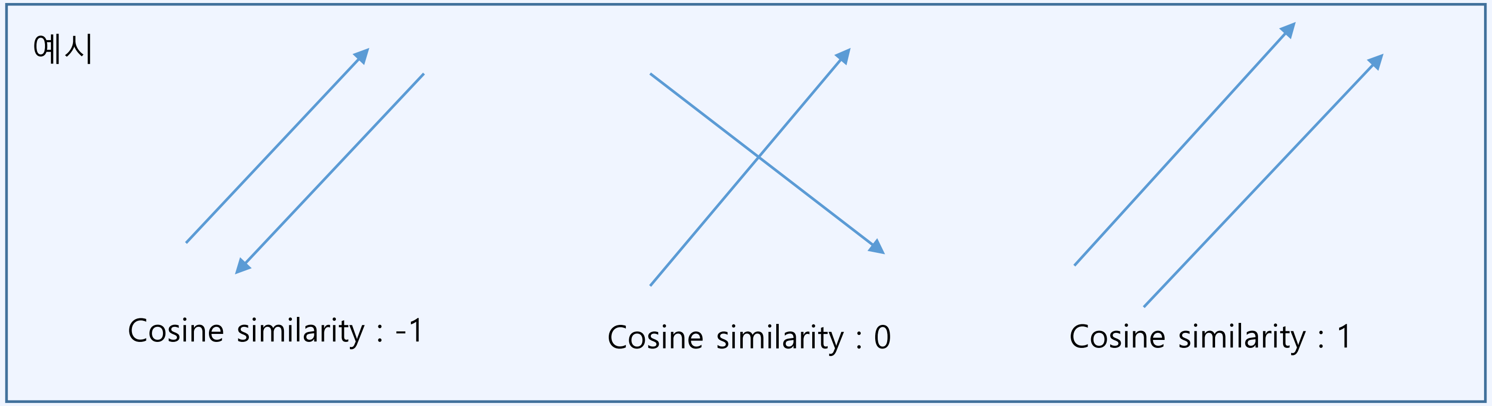
τ

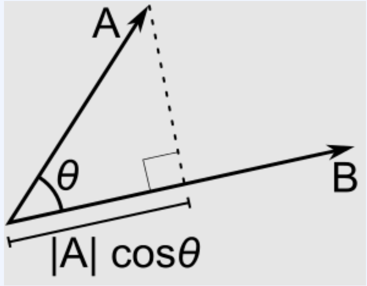
$$softmax(cosine(Q,K)/\tau+b)V$$

Swin Transformer block

- Scaled cosine function

$$Sim(q_i, k_j) = cosine(q_i, k_j)/\tau + B_{ij}$$



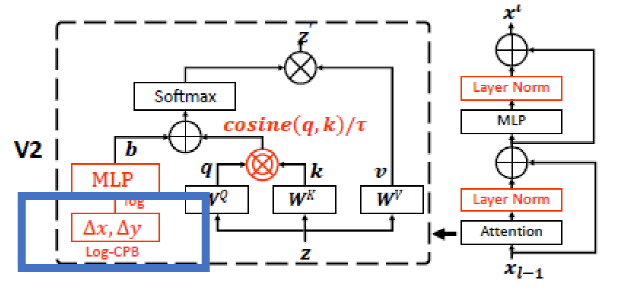


$$\mathrm{A} \cdot B = ||A||\,||B||cos\theta$$

$$\cos(\theta) = \frac{\mathrm{A} \cdot B}{||A||\,||B||}$$

$$Sim(q_i, k_j) = \frac{q_i \cdot k_j}{||q_i||\,||k_j||}/\tau + B_{ij}$$

log-spaced CPB



- Relative coordinates

  Window size(M) = 3

x axis

1	2	3
4	5	6
7	8	9

Δx

	1	2	3	4	5	6	7	8	9
1	0	0	0	-1	-1	-1	-2	-2	-2
2	0	0	0	-1	-1	-1	-2	-2	-2
3	0	0	0	-1	-1	-1	-2	-2	-2
4	1	1	1	0	0	0	-1	-1	-1
5	1	1	1	0	0	0	-1	-1	-1
6	1	1	1	0	0	0	-1	-1	-1
7	2	2	2	1	1	1	0	0	0
8	2	2	2	1	1	1	0	0	0
9	2	2	2	1	1	1	0	0	0

y axis

1	2	3
4	5	6
7	8	9

Δy

	1	2	3	4	5	6	7	8	9
1	0	-1	-2	0	-1	-2	0	-1	-2
2	1	0	-1	1	0	-1	1	0	-1
3	2	1	0	2	1	0	2	1	0
4	0	-1	-2	0	-1	-2	0	-1	-2
5	1	0	-1	1	0	-1	1	0	-1
6	2	1	0	2	1	0	2	1	0
7	0	-1	-2	0	-1	-2	0	-1	-2
8	1	0	-1	1	0	-1	1	0	-1
9	2	1	0	2	1	0	2	1	0

log-spaced CPB

- Log-spaced coordinates :

$$\widehat{\Delta x} = \mathrm{sign}(x) \cdot \log(1 + |\Delta x|),$$
$$\widehat{\Delta y} = \mathrm{sign}(y) \cdot \log(1 + |\Delta y|),$$

$\widehat{\Delta x}$

	1	2	3	4	5	6	7	8	9
1	**0**	**0**	**0**	-0.6931	-0.6931	-0.6931	-1.0986	-1.0986	-1.0986
2	0	0	0	-0.6931	-0.6931	-0.6931	-1.0986	-1.0986	-1.0986
3	0	0	0	-0.6931	-0.6931	-0.6931	-1.0986	-1.0986	-1.0986
4	0.6931	0.6931	0.6931	0	0	0	-0.6931	-0.6931	-0.6931
5	0.6931	0.6931	0.6931	0	0	0	-0.6931	-0.6931	-0.6931
6	0.6931	0.6931	0.6931	0	0	0	-0.6931	-0.6931	-0.6931
7	1.0986	1.0986	1.0986	0.6931	0.6931	0.6931	0	0	0
8	1.0986	1.0986	1.0986	0.6931	0.6931	0.6931	0	0	0
9	1.0986	1.0986	1.0986	0.6931	0.6931	0.6931	0	0	0

$\widehat{\Delta y}$

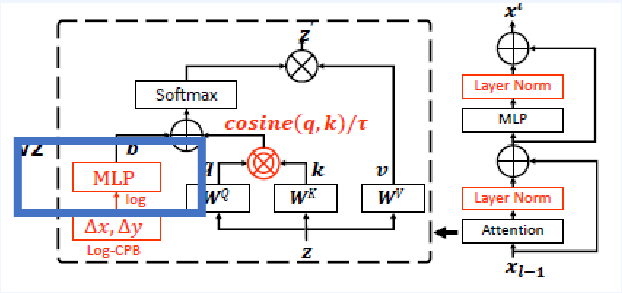
	1	2	3	4	5	6	7	8	9
1	**0**	-0.6931	-1.0986	**0**	-0.6931	-1.0986	**0**	-0.6931	-1.0986
2	0.6931	0	-0.6931	0.6931	0	-0.6931	0.6931	0	-0.6931
3	1.0986	0.6931	0	1.0986	0.6931	0	1.0986	0.6931	0
4	0	-0.6931	-1.0986	0	-0.6931	-1.0986	0	-0.6931	-1.0986
5	0.6931	0	-0.6931	0.6931	0	-0.6931	0.6931	0	-0.6931
6	1.0986	0.6931	0	1.0986	0.6931	0	1.0986	0.6931	0
7	0	-0.6931	-1.0986	0	-0.6931	-1.0986	0	-0.6931	-1.0986
8	0.6931	0	-0.6931	0.6931	0	-0.6931	0.6931	0	-0.6931
9	1.0986	0.6931	0	1.0986	0.6931	0	1.0986	0.6931	0

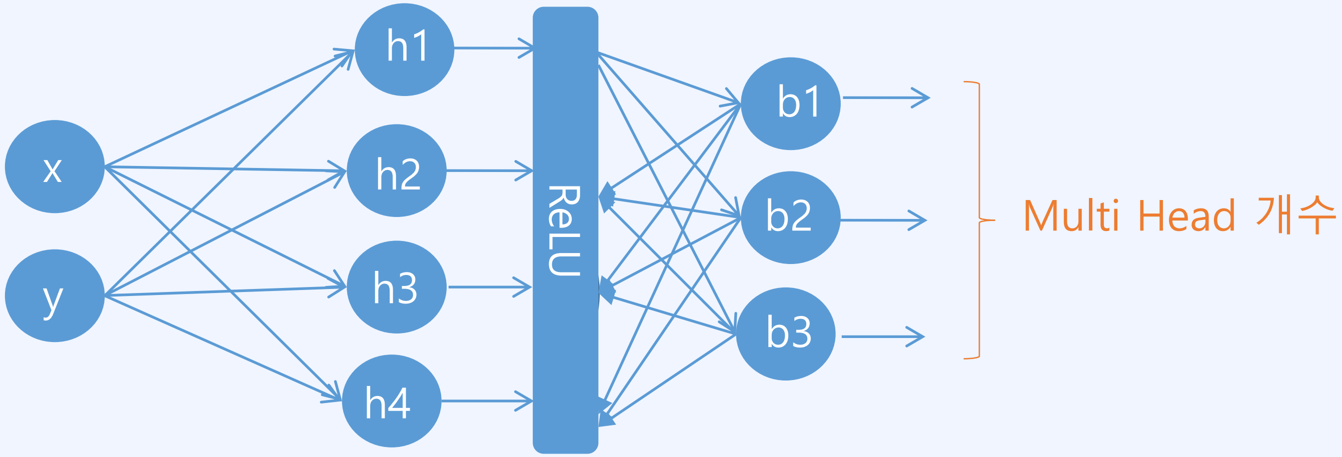
log-spaced CPB

- Continuous relative position bias

$$B(\Delta x, \Delta y) = \mathcal{G}(\Delta x, \Delta y),$$

Meta Network(MLP)





Architecture Variants

- Architecture hyper-parameters

Swin-T: $C = 96$, layer numbers = $\{2, 2, 6, 2\}$

Swin-S: $C = 96$, layer numbers =$\{2, 2, 18, 2\}$

Swin-B: $C = 128$, layer numbers =$\{2, 2, 18, 2\}$

Swin-L: $C = 192$, layer numbers =$\{2, 2, 18, 2\}$

SwinV2-H: $C = 352$, #. block = $\{2, 2, 18, 2\}$

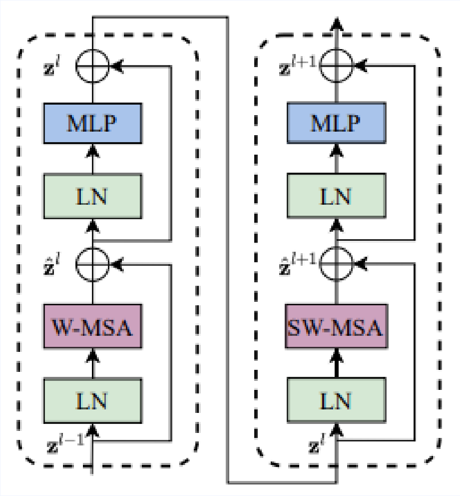
SwinV2-G: $C = 512$, #. block = $\{2, 2, 42, 4\}$

** C는 첫 번째 Stage의 hidden layer의 channel 개수*

Summary

- Hierarchical representation을 학습
  - Object Detection task 수행
  - Backbone으로 사용

Two Swin Transformer blocks



Architecture