NIPS 2017

Attention Is All You Need

ICLR 2021

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Micosoft 2021

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

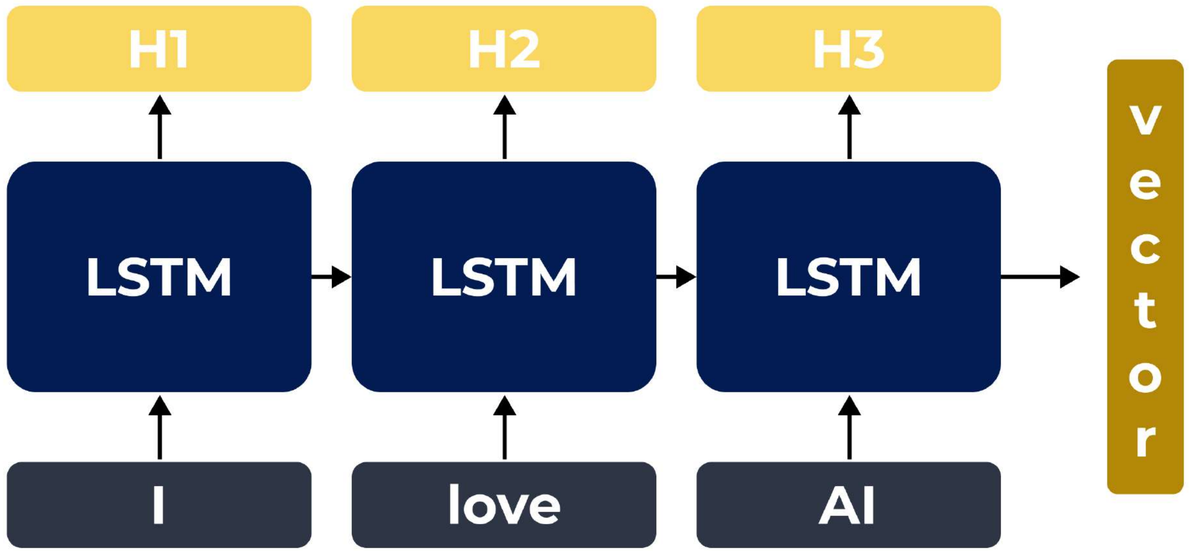
Minsu koh

Transformer

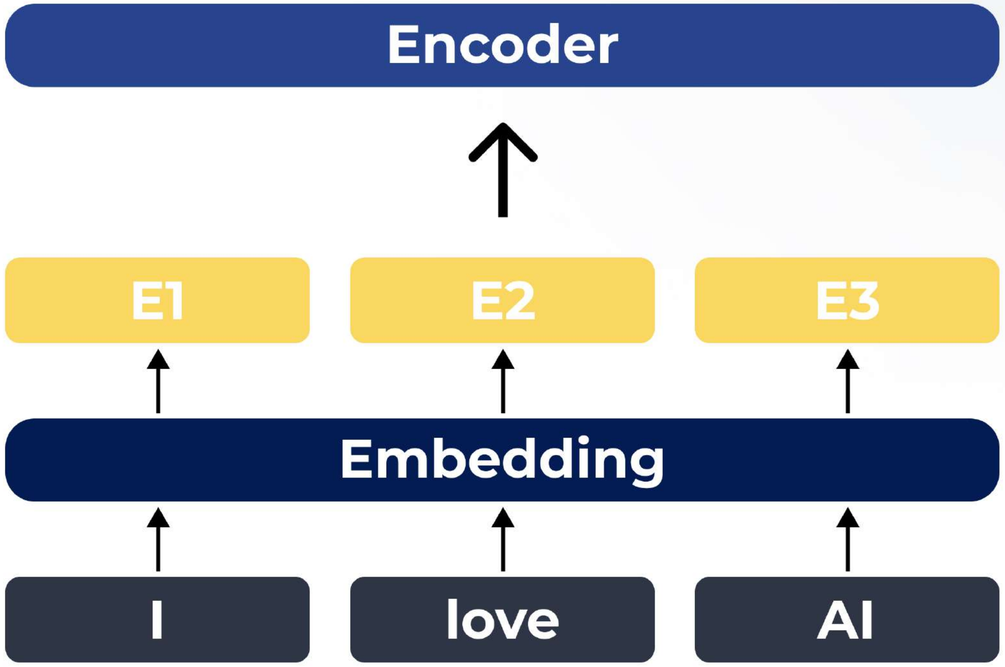
Attention Is All You Need

Why we need Transformer?

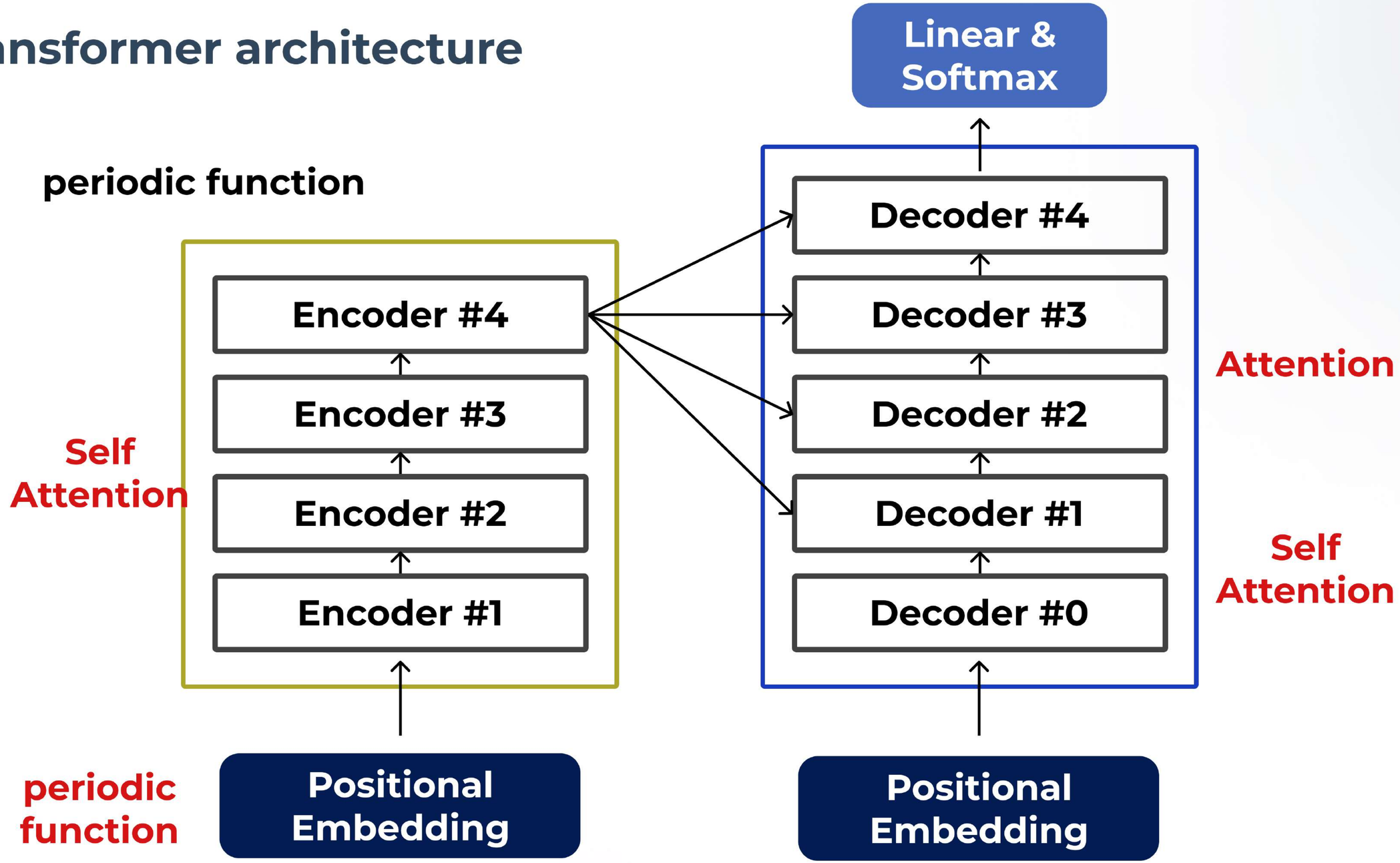
Seq2Seq



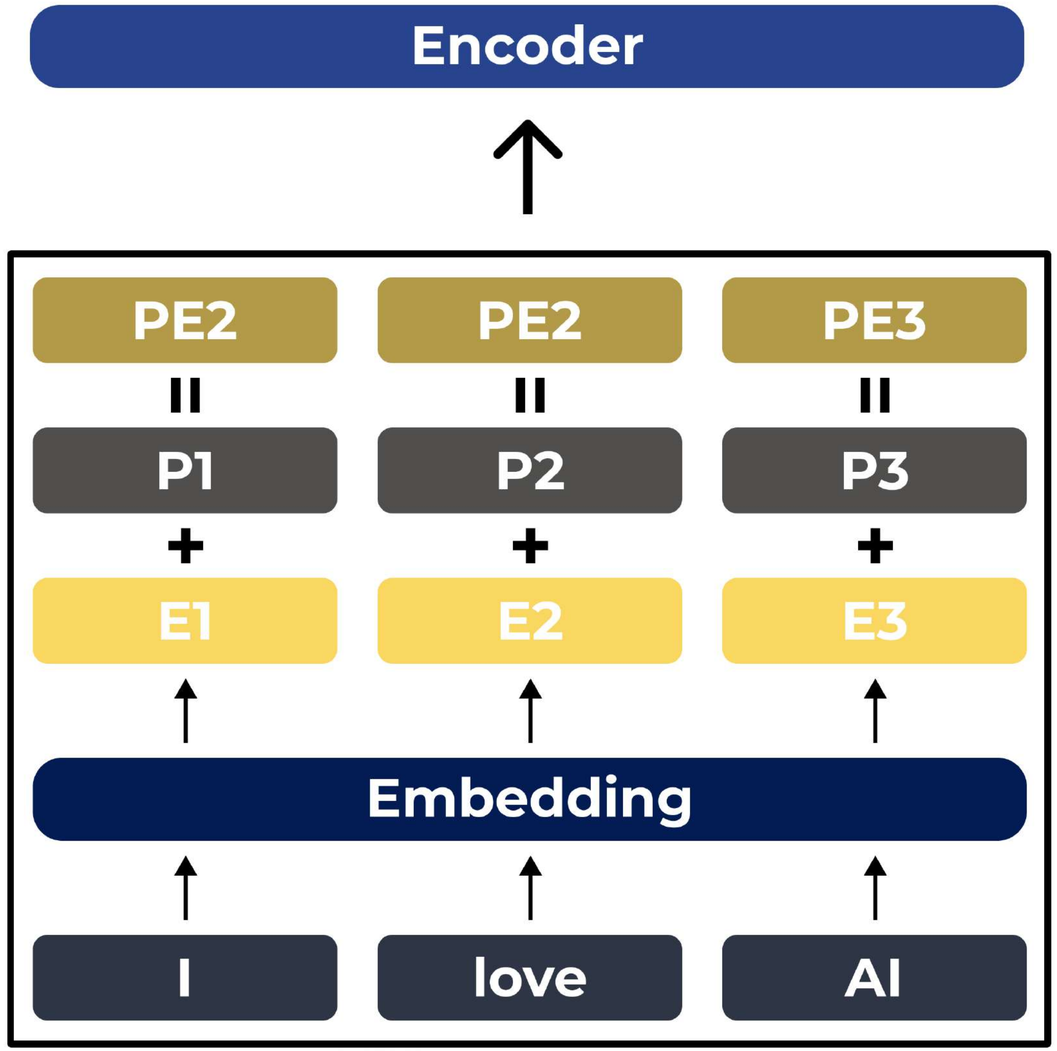
Transformer



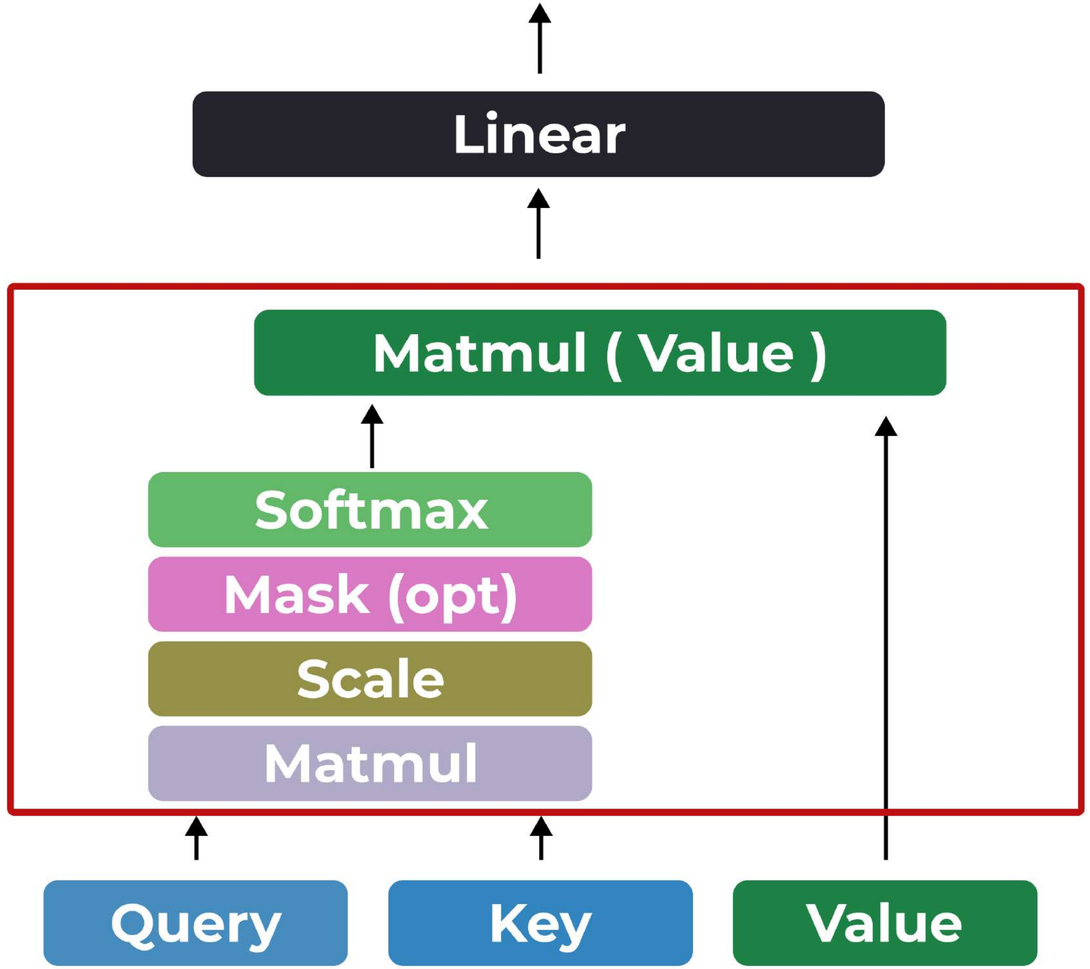
Transformer architecture



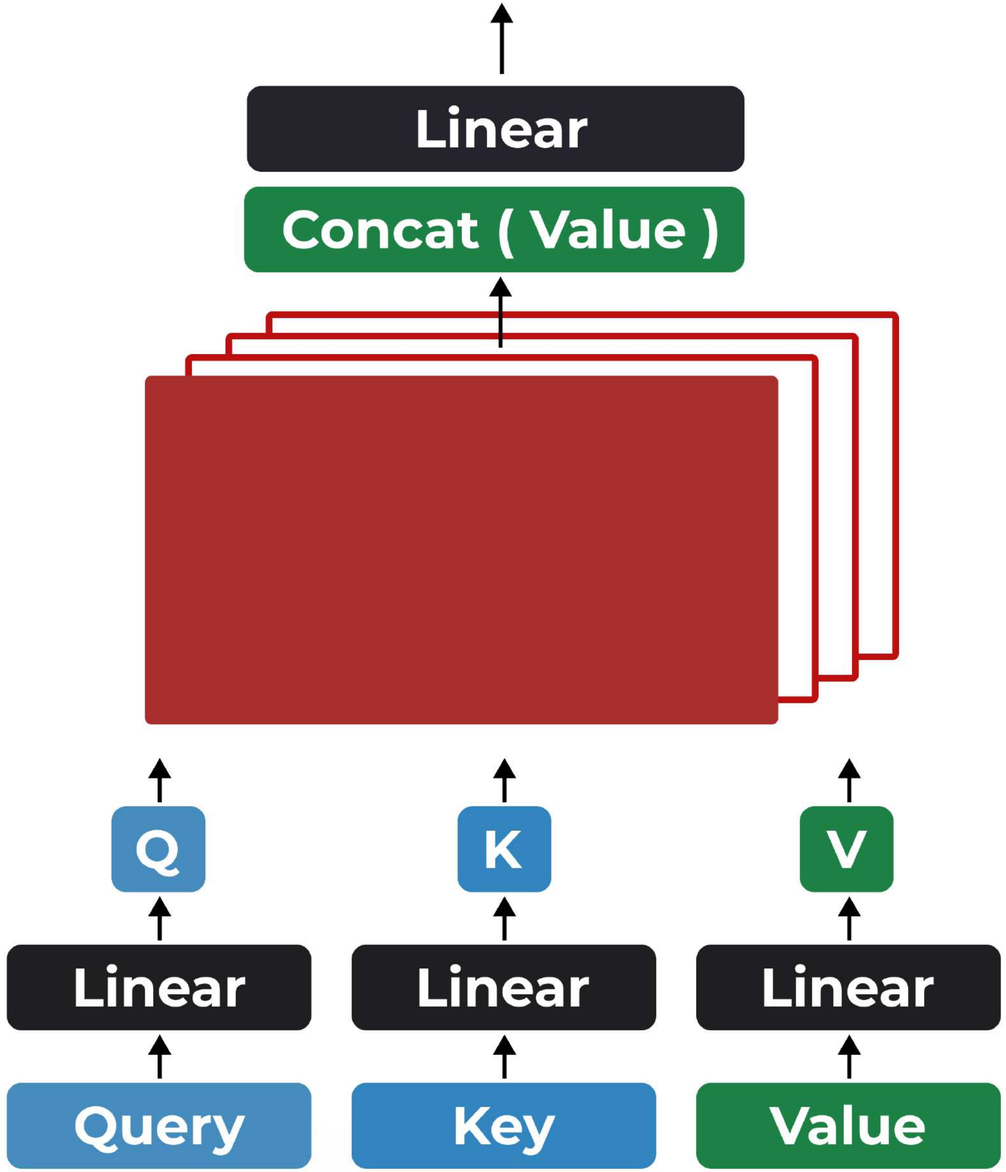
Positional Embedding



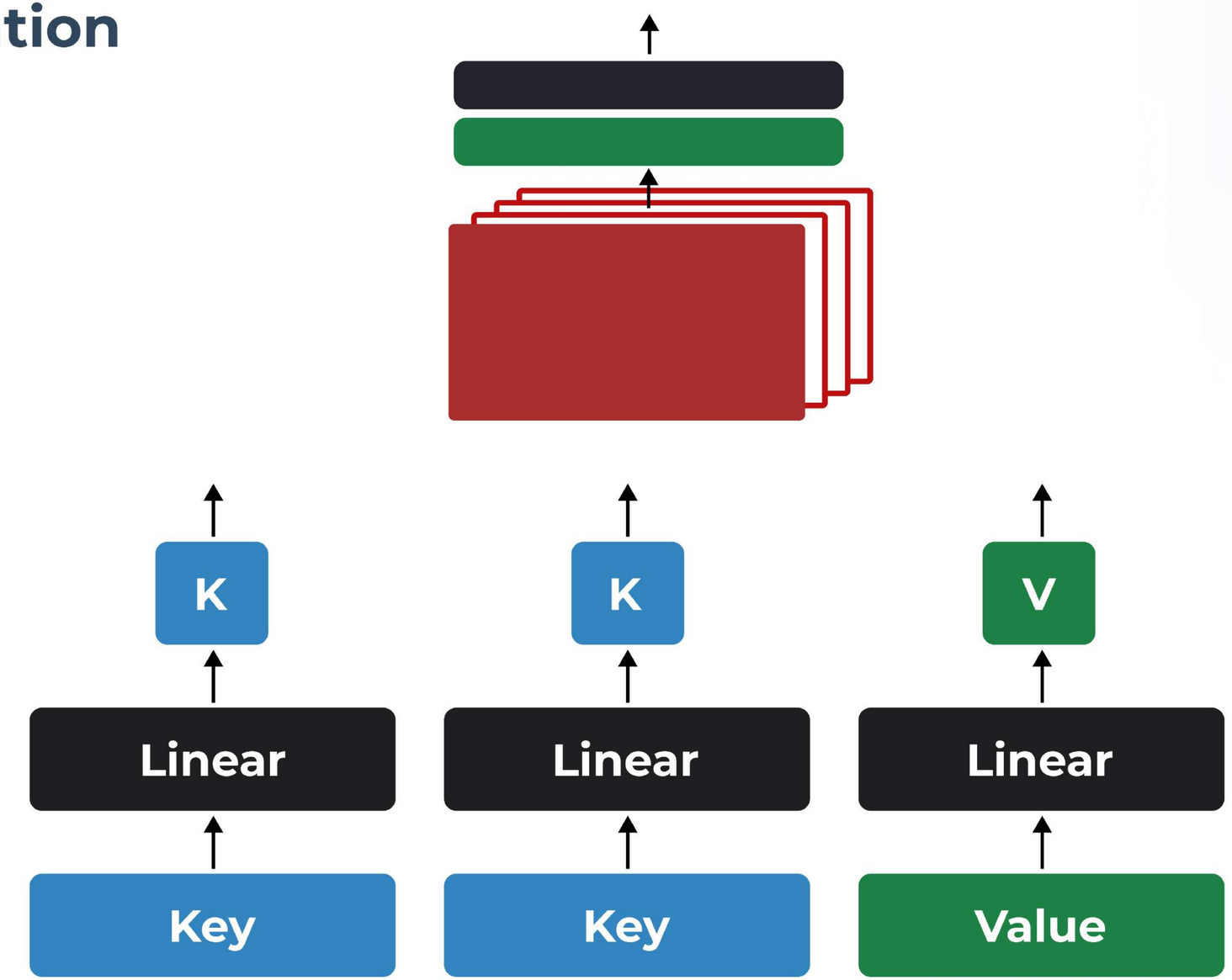
Attention (One head Attention)



Attention (Multi head Attention)

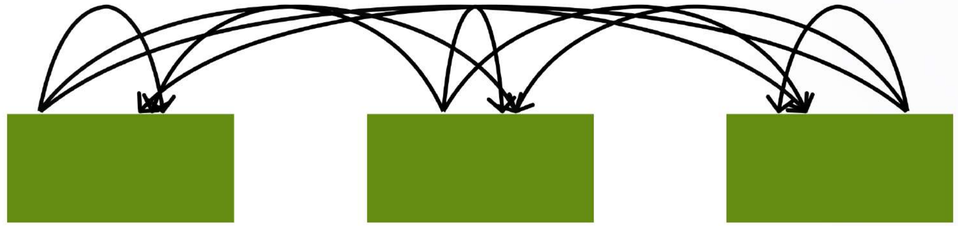


Self Attention

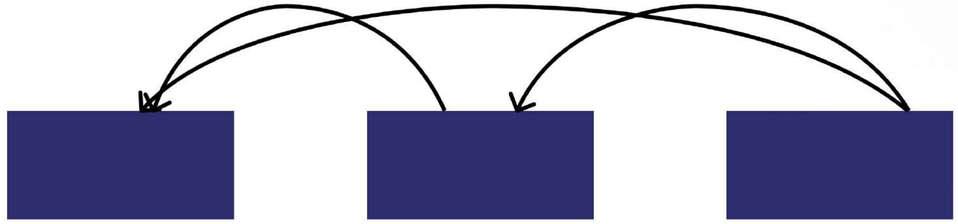


Attention type

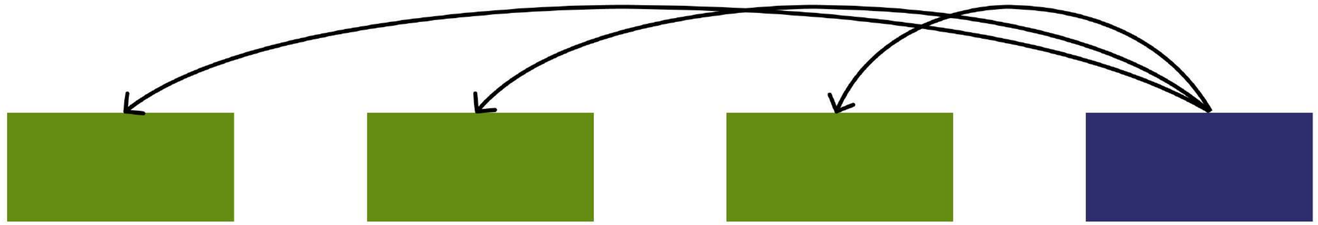
Encoder Self Attention



Masked Decoder Self Attention



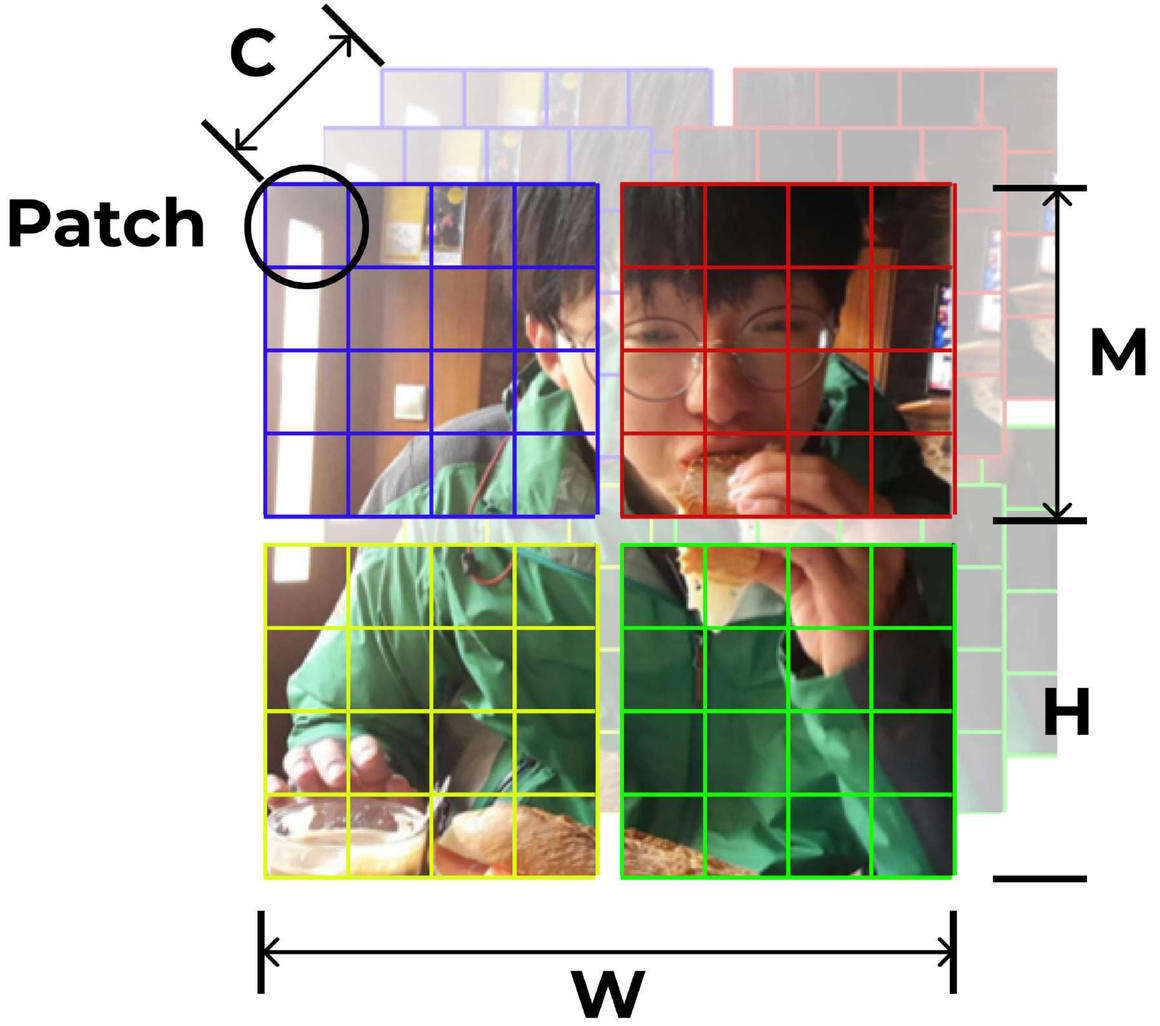
Encoder-Decoder Attention



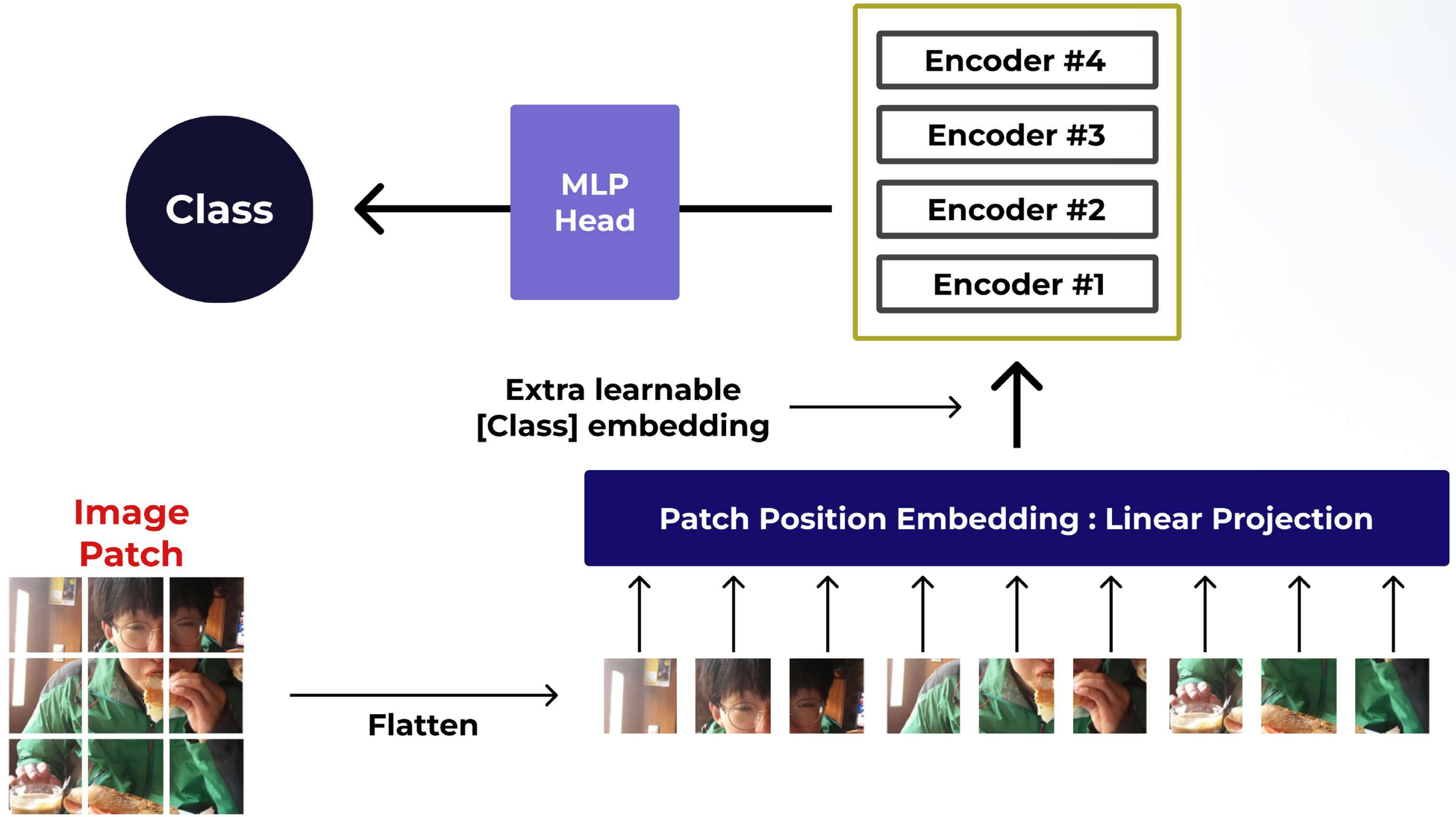
Vision Transformer

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Notice



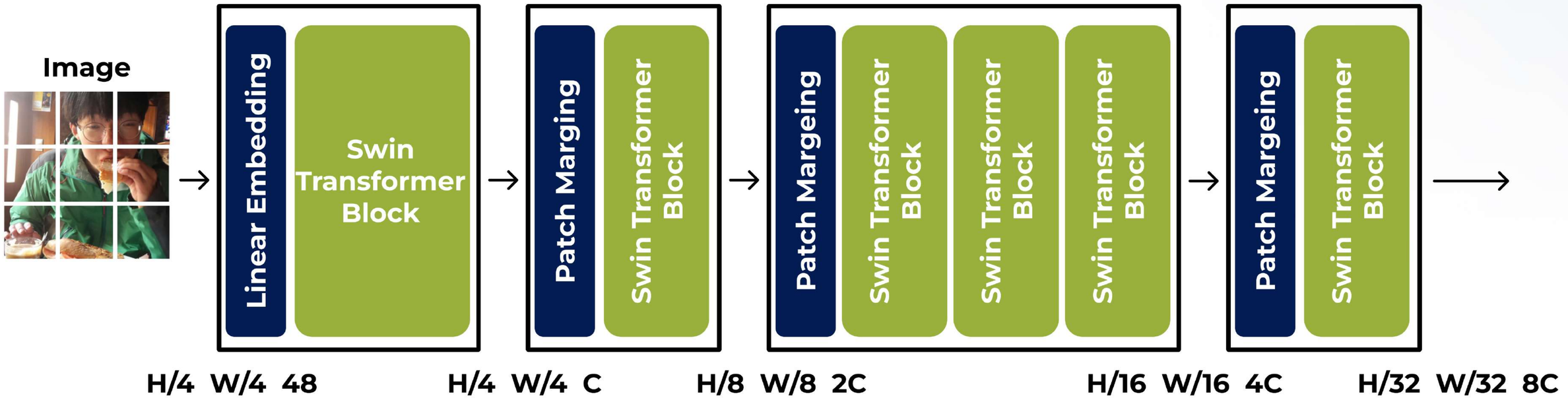
Vision Transformer architecture



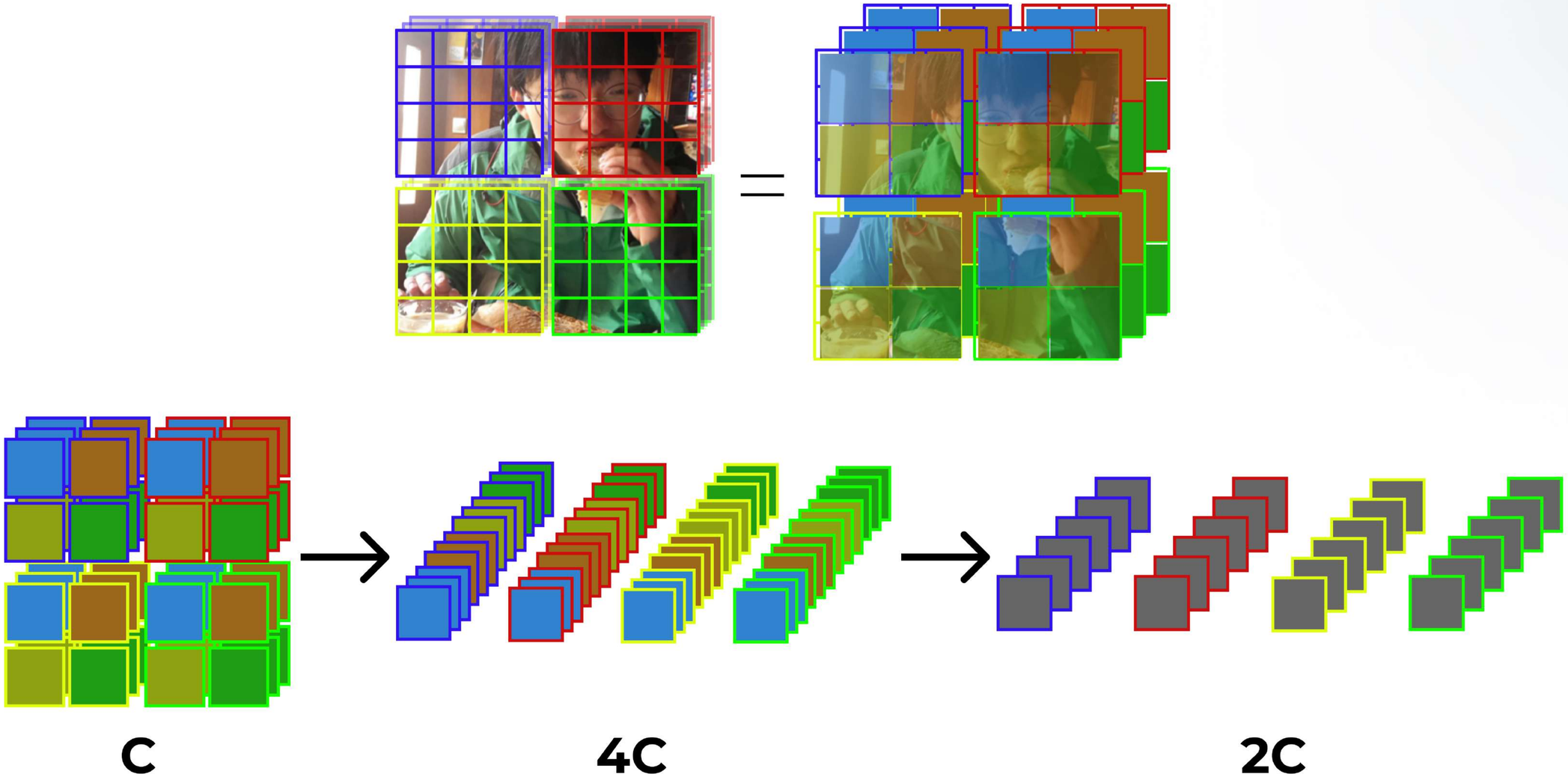
Swin Transformer

Swin Transformer:
Hierarchical Vision Transformer using Shifted Windows

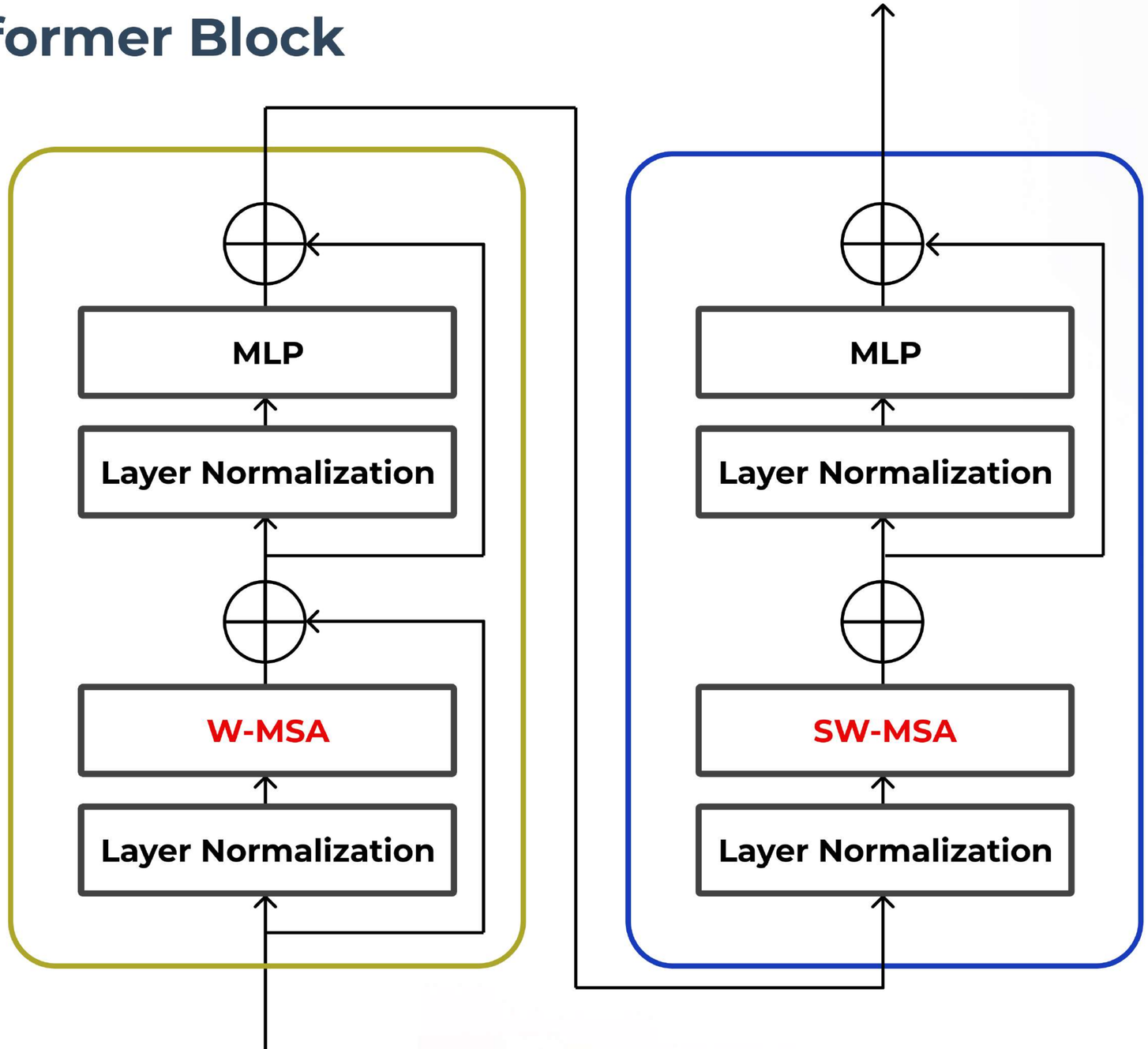
Swin Transformer achitecture



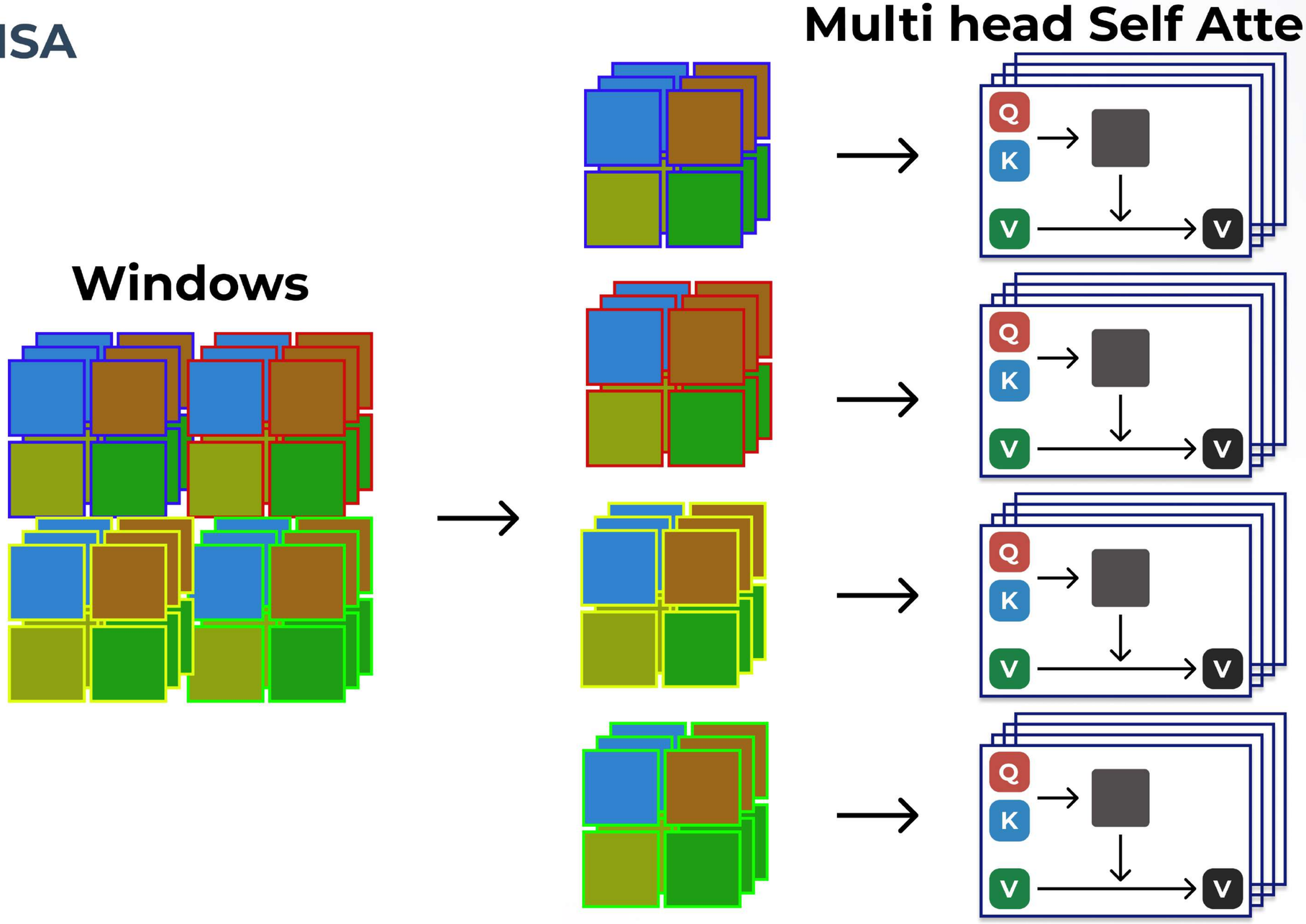
Patch Marging



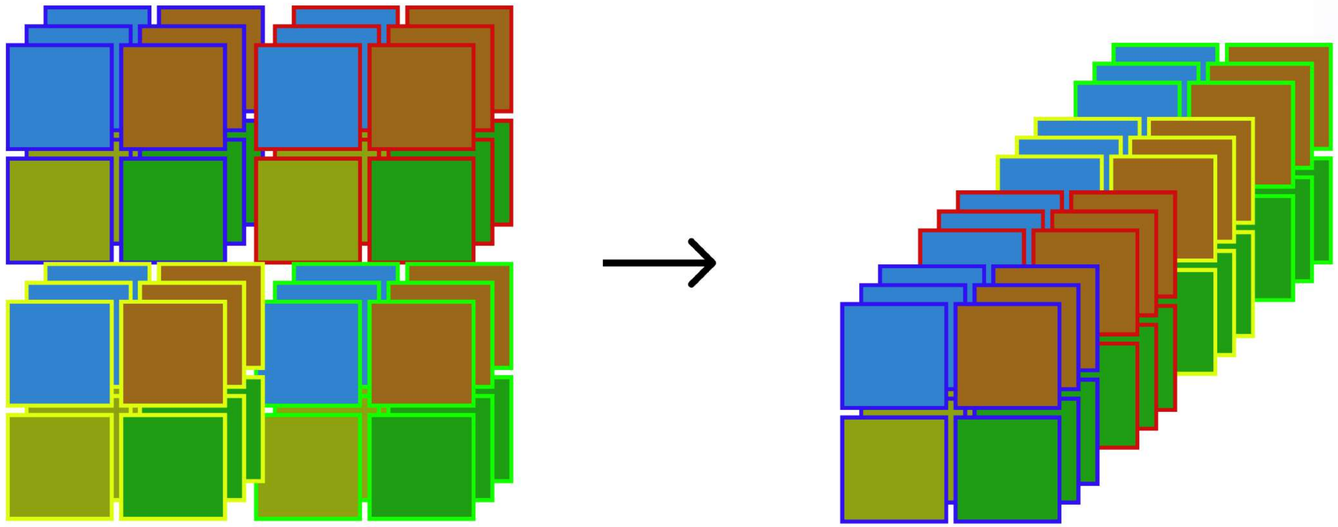
Swin Transformer Block



W-MSA

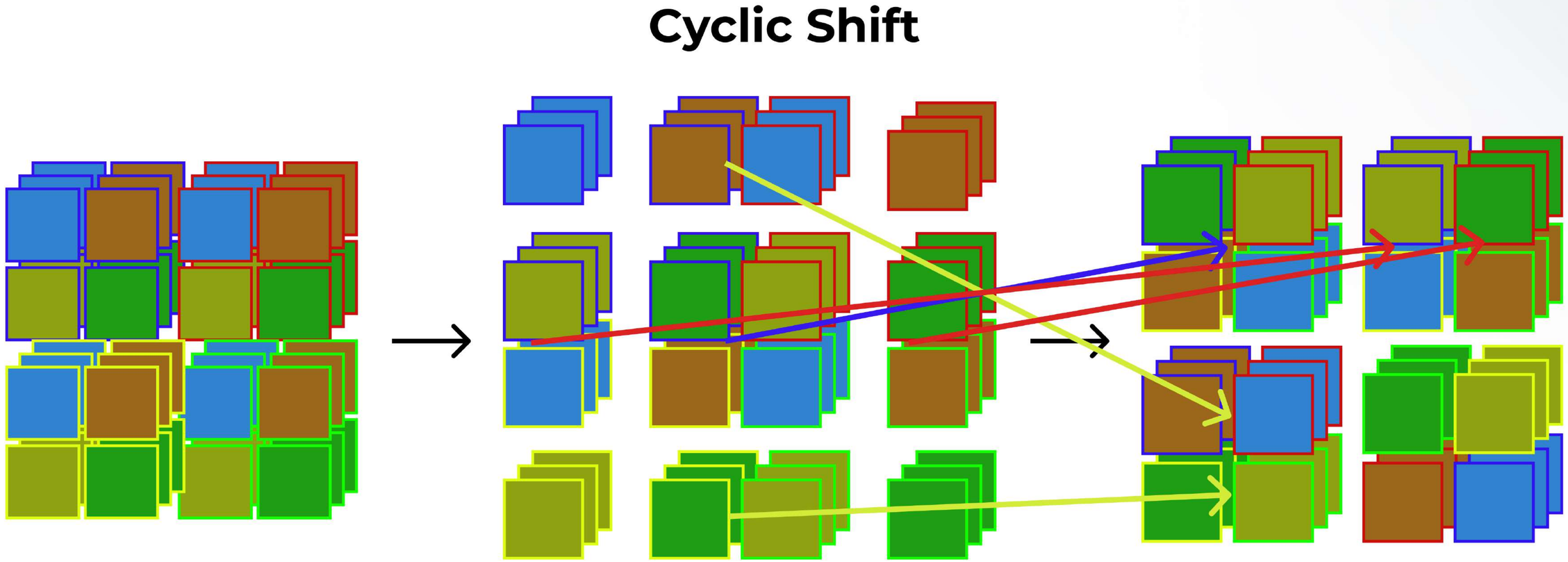


SW-MSA - (Efficient Batch Computation)



$B \times N_h \times N_w \times C$ $\quad$ $nB \times M \times M \times C$

SW-MSA - (Cycliv Shift)



SW-MSA - (Relative Position Bias)

if window_size = 3

$$\text{Attention}(Q, K, V) = \text{SoftMax}(QK^T/\sqrt{d} + B)V,$$

x-axis Matrix

1	2	3
4	5	6
7	8	9

0	0	0	-1	-1	-1	-2	-2	-2
0	0	0	-1	-1	-1	-2	-2	-2
0	0	0	-1	-1	-1	-2	-2	-2
1	1	1	0	0	0	-1	-1	-1
1	1	1	0	0	0	-1	-1	-1
1	1	1	0	0	0	-1	-1	-1
2	2	2	1	1	1	0	0	0
2	2	2	1	1	1	0	0	0
2	2	2	1	1	1	0	0	0

+= window_size - 1

***= 2*window_size - 1**

y-axis Matrix

1	2	3
4	5	6
7	8	9

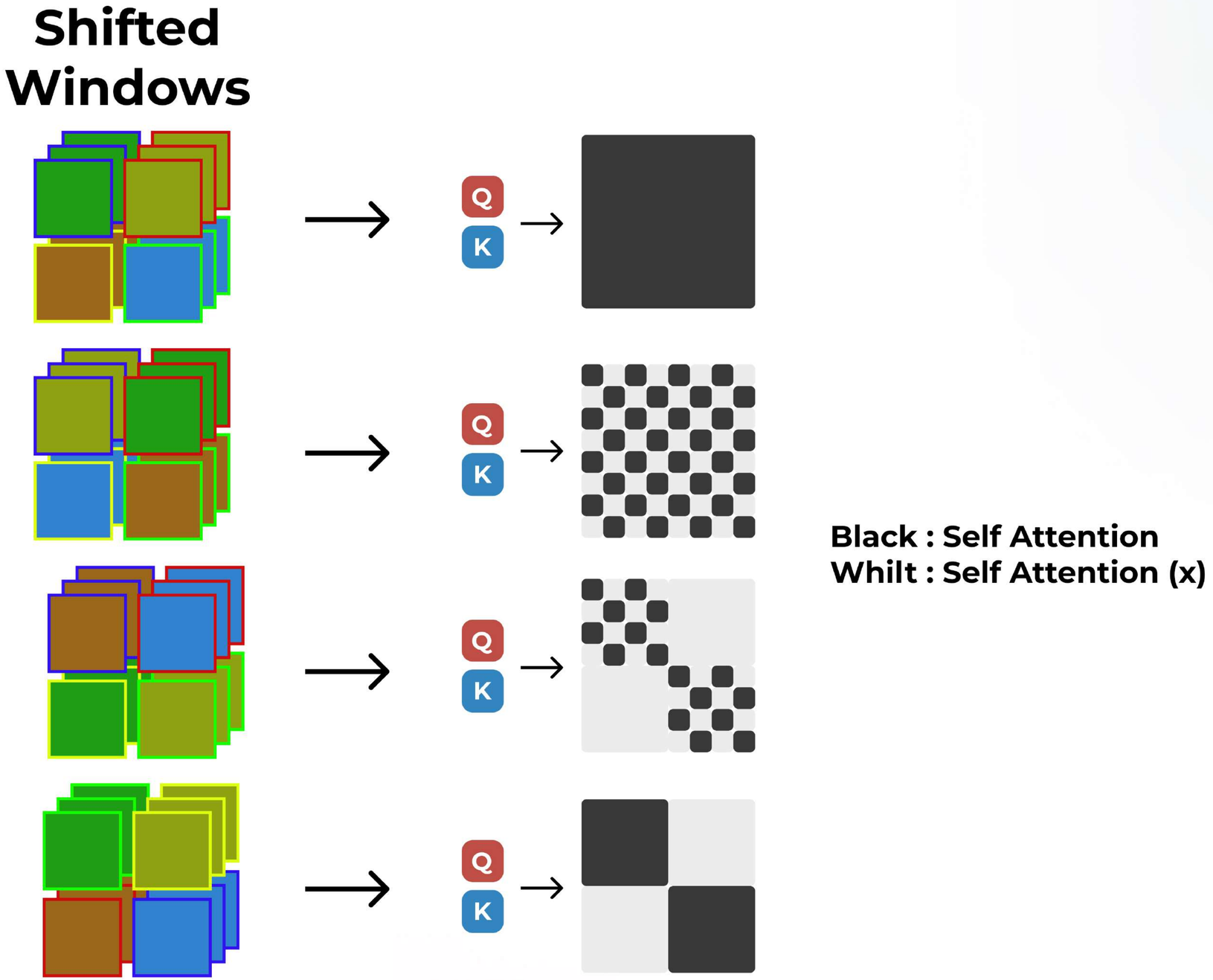
0	-1	-2	0	-1	-2	0	-1	-2
1	0	-1	1	0	-1	1	0	-1
2	1	0	2	1	0	2	1	0
0	-1	-2	0	-1	-2	0	-1	-2
1	0	-1	1	0	-1	1	0	-1
2	1	0	2	1	0	2	1	0
0	-1	-2	0	-1	-2	0	-1	-2
1	0	-1	1	0	-1	1	0	-1
2	1	0	2	1	0	2	1	0

+= window_size - 1

Relative Position Bias

12	11	10	7	6	5	2	1	0
13	12	11	8	7	6	3	2	1
14	13	12	9	8	7	4	3	2
17	16	15	12	11	10	7	6	5
18	17	16	13	12	11	8	7	6
19	18	17	14	13	12	9	8	7
22	21	20	17	16	15	12	11	10
23	22	21	18	17	16	13	12	11
24	23	22	19	18	17	14	13	12

SW-MSA


Shifted Windows
Q
K
Q
K
Q
K
Q
K
Black : Self Attention
Whilt : Self Attention (x)

Thank you

Attention Is All You Need

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Minsu koh