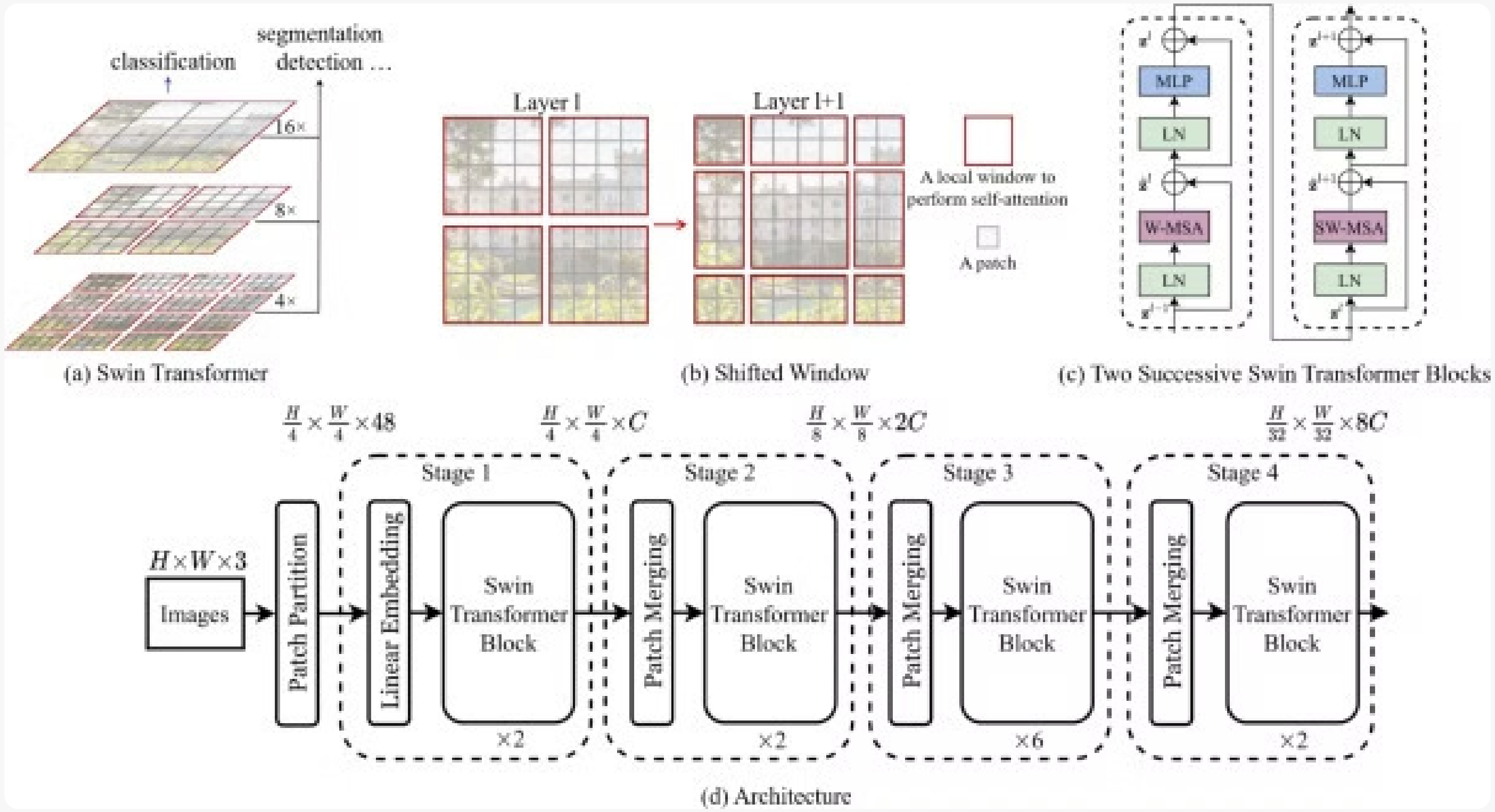
Introduce Swin Transformer



Swin Transformer (Shifted Window Transformer) is a novel model for computer vision tasks. This paper proposes using the Swin Transformer as the backbone for computer vision tasks such as object detection, segmentation, and others.

작성자: /AI·소프트웨어학부(인공지능전공) 정승민

The emergence of the Transformer in computer vision

1

Limitation of the CNN

CNN is very powerful model for understanding local information. However there is a difficulty in capturing the overall context

2

Emergence of the Transformer

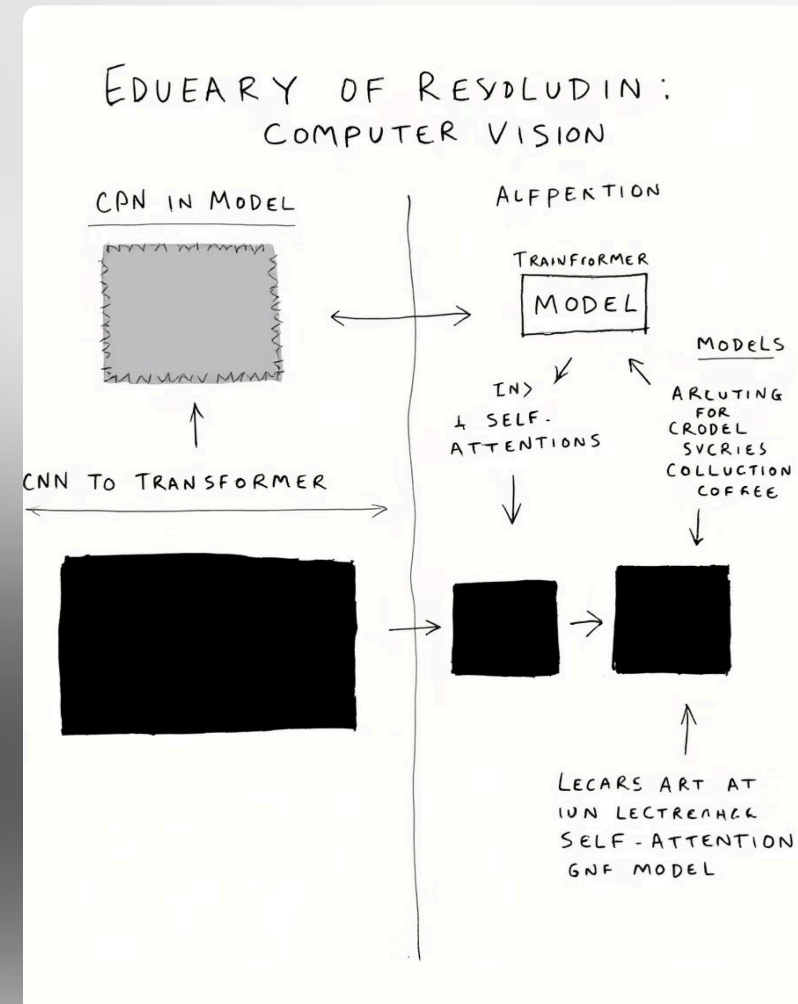
The Transformer has succeed in NLP by using self-attention mechanism that is effective at learning global context.

Moreover, the transformer has been adapted for vision task

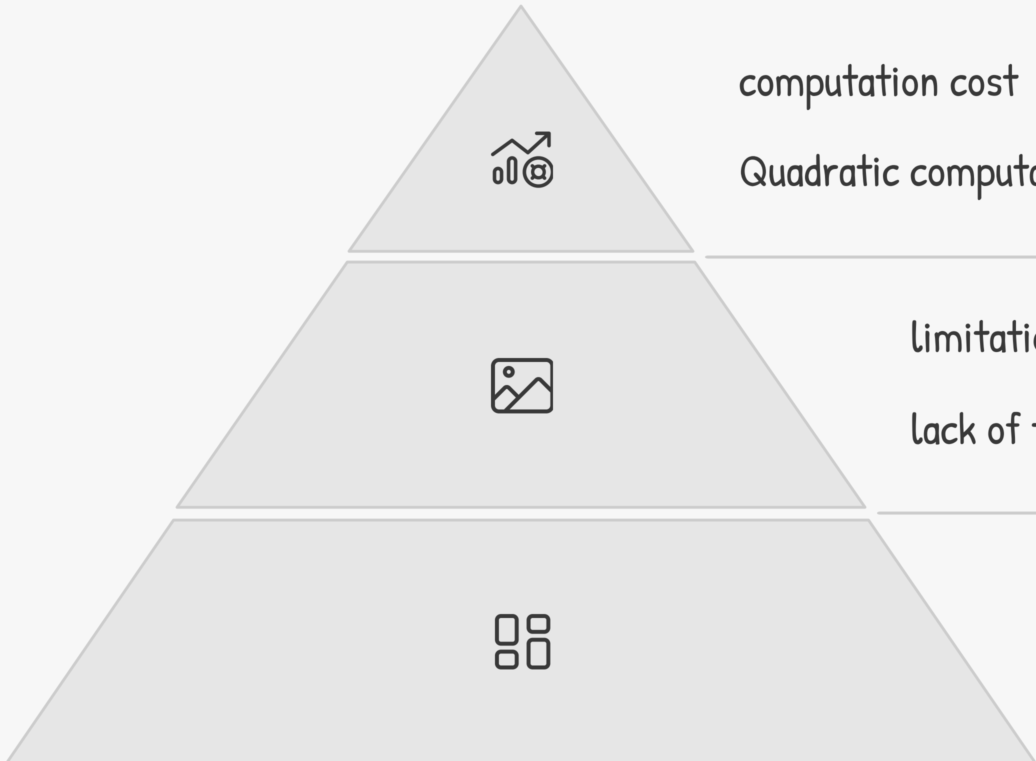
3

Emergence of the ViT and limitation

ViT is a successful model that adapts the transformer for images. It is a powerful model in image classification. However, if the image resolution increase, computational cost increase quadratically.



Challenges of the Transformer (ViT)



computation cost

Quadratic computational complexity with respect to image size

limitation of resolution

lack of the flexibility due to the fixed resolution

limitation of multi-scale capability

Difficulty in handling objects of various size

Quadratic computational complexity with respect to image resolution is a critical limitation of ViT, making it challenging to use in real-world applications. Because ViT's self-attention compute the relationships between each patch and all other patches globally

ViT is implemented with fixed images resolution, so it lacks the flexibility to handle varying image resolutions.

Also, since ViT lacks a hierarchical structure, it has d to detect semantic patterns and object of varying scales.

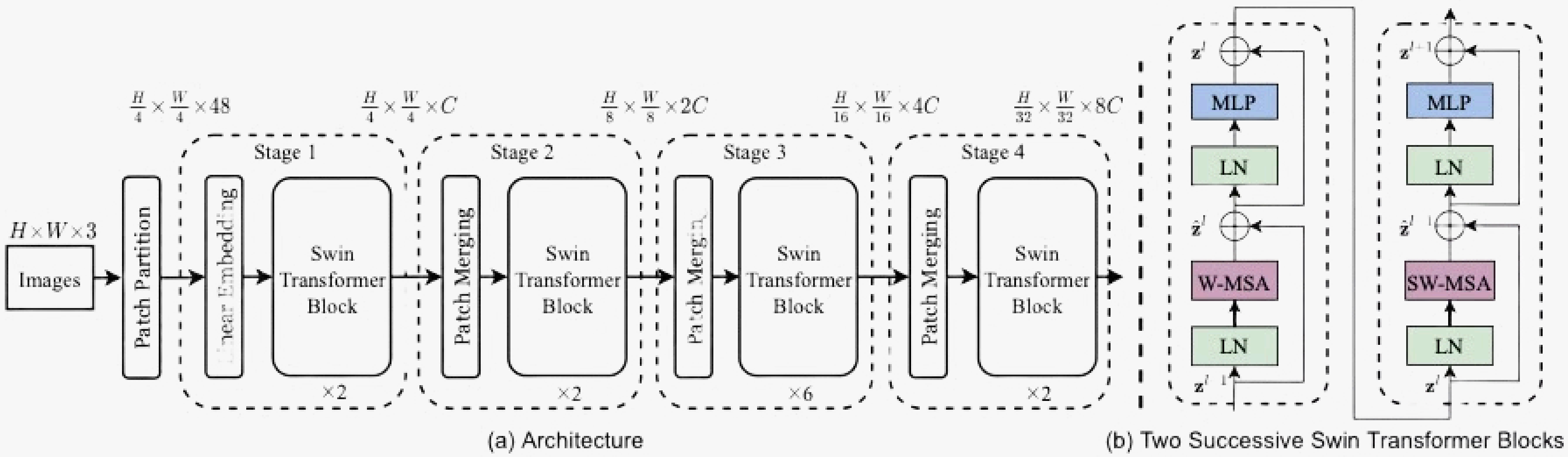
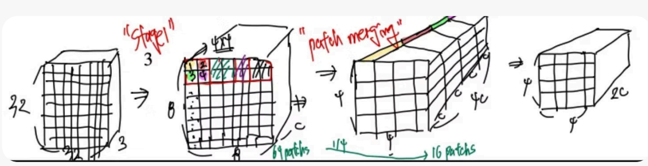


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

Swin Transformer's main idea

Hierarchical structure



self-attention based shifted window

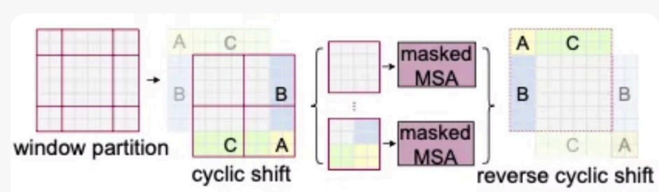
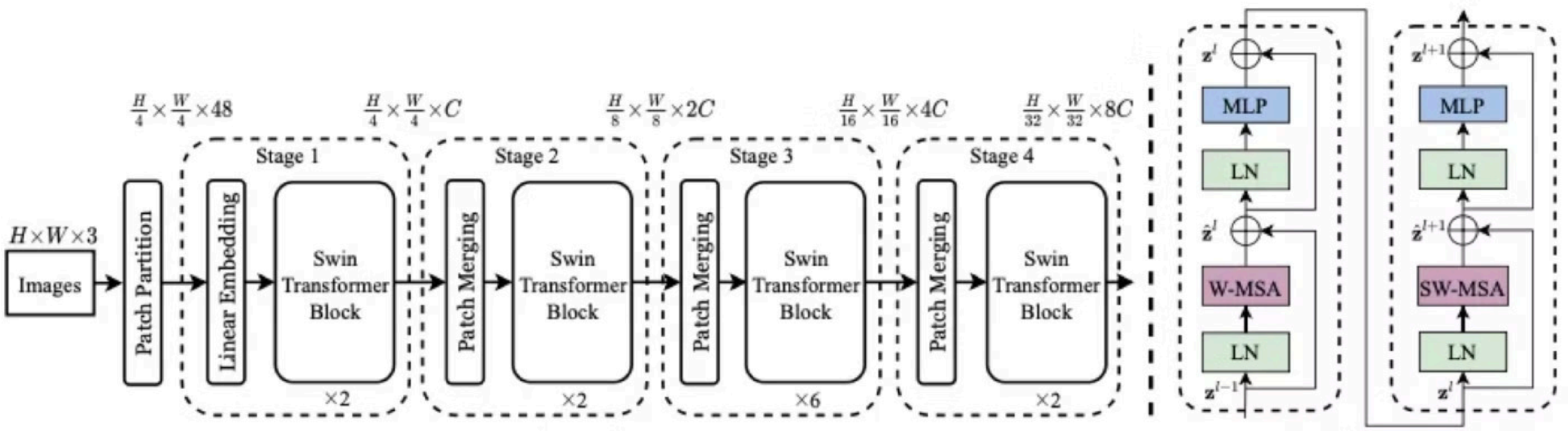


Figure 4. Illustration of an efficient batch computation approach for self-attention in shifted window partitioning.

Linear complexity respect to image resolution

$$\Omega(\mathrm{MSA}) = 4hwC^2 + 2(hw)^2C, \quad (1)$$
$$\Omega(\mathrm{W\text{-}MSA}) = 4hwC^2 + 2M^2hwC, \quad (2)$$

Swin Transformer's architecture



(a) Architecture (b) Two Successive Swin Transformer Blocks

Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

Initial process

The input image is divided into small patches (typically 4x4) and converted into C-dimensional feature vectors via a linear embedding layer. This process is similar to ViT, but the subsequent processing differs.

4 step hierarchical structure

Swin Transformer features a 4-stage hierarchical architecture where resolution decreases and channel dimensions increase progressively, capturing multi-scale features similar to CNNs.

patch merging laeyr

Between stages, adjacent patches are merged, reducing the token count while doubling the channel dimension and halving the resolution at each stage.

Swin Transformer block

Each stage contains multiple Swin Transformer blocks, which alternate between window-based and shifted window self-attention to perform the core computations.

Composition of Swin Transformer Block

1

Window-based Multi-head Self Attention (W-MSA)

Each Swin Transformer block first applies window-based multi-head self-attention (W-MSA). This mechanism splits the input feature map into non-overlapping windows and computes self-attention within each window, which significantly reduces computational complexity.

2

Shifted-Window Multi-head Self-Attention (SW-MSA)

In the subsequent Swin Transformer block, shifted window multi-head self-attention (SW-MSA) is employed. By shifting the positions of the windows, it enables information exchange between patches that were previously in different windows, enhancing connectivity across windows.

3

MLP and Normalization

After each attention layer, an MLP (Multi-Layer Perceptron) block consisting of two linear layers is applied. Additionally, LayerNorm normalization is used before each sub-block, and a residual connection is added after each sub-block. This design follows conventional practices.

** MLP employs GLEU as activation function

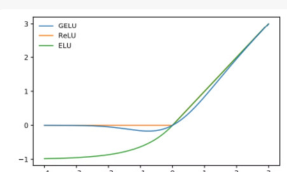
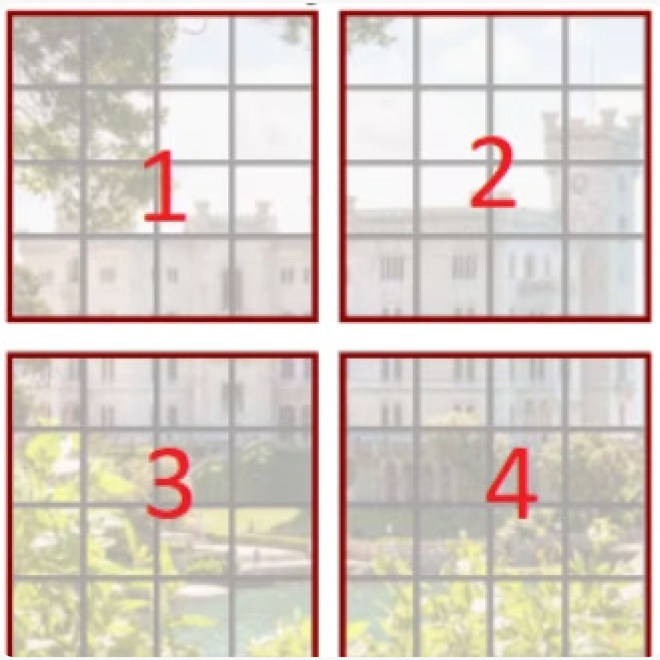


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).

Window-based Self-attention



Non-overlapped window

→ Can focus local information

Linear complexity

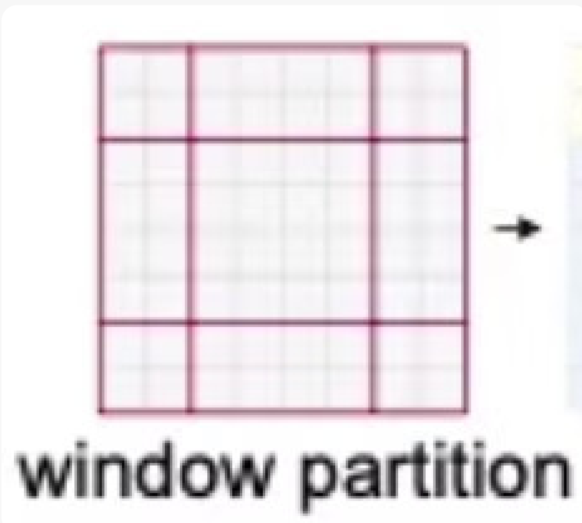
→ $O(N^2) \rightarrow O(N)$

Efficient parallel compute

→ Can compute parallel.

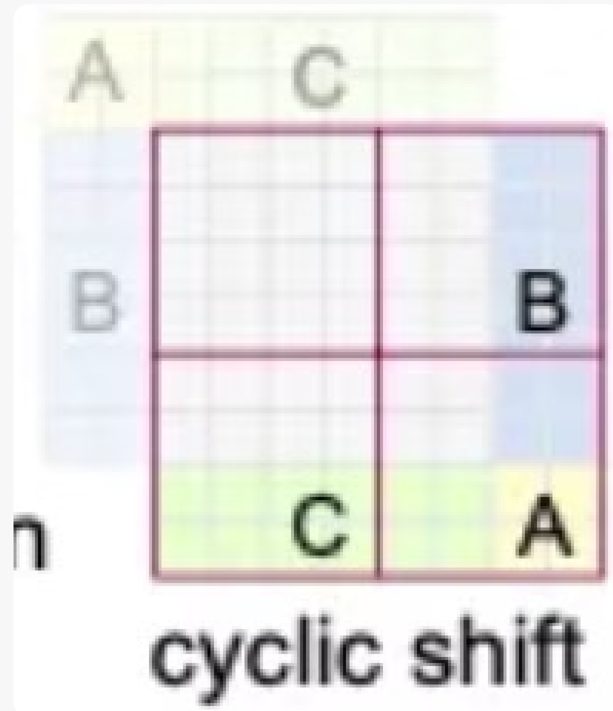
Shifted Window Self-attention

Shifted Window Mechanism

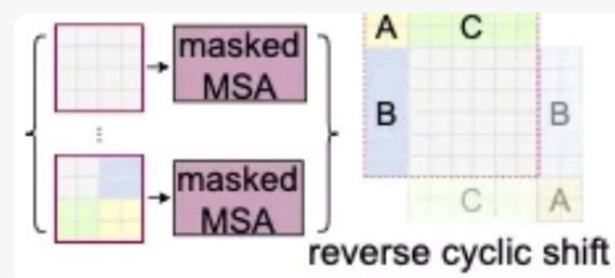


Cross-Window Connectivity

Efficient Cyclic Shift



Masking Technique.



Shifted Window Self-Attention is the most innovative part of the Swin Transformer. In the first block, regular window partitioning is used, and in the subsequent block, the windows are shifted by $(\lfloor M/2 \rfloor, \lfloor M/2 \rfloor)$ pixels. This shift enables information exchange between patches that were previously in different windows.

Although this shifted window mechanism greatly enhances connectivity between patches, it results in the creation of more windows (For example above Cross-Window connectivity's right image has 9 window). To efficiently address this, the Swin Transformer employs a cyclic shift technique to maintain the original number of windows..

Performance and Advantage of the Swin Transformer

87.3%

Object detection accuracy in COCO dataset (AP)

53.5%

semantic segmentation

ADE20K dataset (mIOU)

4x

Computation efficient

86.4%

ImageNet accuracy

ImageNet-1K dataset (classification)

Object Detection

(a) Various frameworks

Method	Backbone	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	#param.	FLOPs	FPS
Cascade	R-50	46.3	64.3	50.5	82M	739G	18.0
Mask R-CNN	Swin-T	**50.5**	**69.3**	**54.9**	86M	745G	15.3
ATSS	R-50	43.5	61.9	47.0	32M	205G	28.3
	Swin-T	**47.2**	**66.5**	**51.3**	36M	215G	22.3
RepPointsV2	R-50	46.5	64.6	50.3	42M	274G	13.6
	Swin-T	**50.0**	**68.5**	**54.2**	45M	283G	12.0
Sparse	R-50	44.5	63.4	48.2	106M	166G	21.0
R-CNN	Swin-T	**47.9**	**67.3**	**52.3**	110M	172G	18.4

(b) Various backbones w. Cascade Mask R-CNN

	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	param	FLOPs	FPS
DeiT-S†	48.0	67.2	51.7	41.4	64.2	44.3	80M	889G	10.4
R50	46.3	64.3	50.5	40.1	61.7	43.4	82M	739G	18.0
Swin-T	**50.5**	**69.3**	**54.9**	**43.7**	**66.6**	**47.1**	86M	745G	15.3
X101-32	48.1	66.5	52.4	41.6	63.9	45.2	101M	819G	12.8
Swin-S	**51.8**	**70.4**	**56.3**	**44.7**	**67.9**	**48.5**	107M	838G	12.0
X101-64	48.3	66.4	52.3	41.7	64.0	45.1	140M	972G	10.4
Swin-B	**51.9**	**70.9**	**56.5**	**45.0**	**68.4**	**48.7**	145M	982G	11.6

Semantic segmentation

ADE20K Method	Backbone	val mIoU	test score	#param.	FLOPs	FPS
DANet [23]	ResNet-101	45.2	-	69M	1119G	15.2
DLab.v3+ [11]	ResNet-101	44.1	-	63M	1021G	16.0
ACNet [24]	ResNet-101	45.9	38.5	-		
DNL [71]	ResNet-101	46.0	56.2	69M	1249G	14.8
OCRNet [73]	ResNet-101	45.3	56.0	56M	923G	19.3
UperNet [69]	ResNet-101	44.9	-	86M	1029G	20.1
OCRNet [73]	HRNet-w48	45.7	-	71M	664G	12.5
DLab.v3+ [11]	ResNeSt-101	46.9	55.1	66M	1051G	11.9
DLab.v3+ [11]	ResNeSt-200	48.4	-	88M	1381G	8.1
SETR [81]	T-Large‡	50.3	61.7	308M	-	-
UperNet	DeiT-S†	44.0	-	52M	1099G	16.2
UperNet	Swin-T	46.1	-	60M	945G	18.5
UperNet	Swin-S	49.3	-	81M	1038G	15.2
UperNet	Swin-B‡	51.6	-	121M	1841G	8.7
UperNet	Swin-L‡	**53.5**	**62.8**	234M	3230G	6.2

Table 3. Results of semantic segmentation on the ADE20K val and test set. † indicates additional deconvolution layers are used to produce hierarchical feature maps. ‡ indicates that the model is pre-trained on ImageNet-22K.

Conclusion and future Research Direction

Efficient Vision Backbone

A single architecture applicable to a range of vision tasks.

Research Direction

Exploration of larger models and diverse applications.

Swin Transformer provides an efficient and flexible backbone network for computer vision through its innovative combination of a hierarchical architecture and shifted window-based self-attention. This model has demonstrated outstanding performance in various vision tasks, including image classification, object detection, and semantic segmentation.

In particular, by leveraging linear computational complexity and hierarchical feature representation, it effectively overcomes the limitations of conventional Vision Transformers—a critical advantage for real-world applications. The success of Swin Transformer serves as a compelling example of the potential of Transformer-based models in computer vision.

Future research directions include pre-training with larger models and datasets, processing higher-resolution images, expanding into spatiotemporal tasks such as video understanding, and integrating self-supervised learning techniques. Swin Transformer is poised to provide a robust foundation for the future of computer vision.