Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu[†*] Yutong Lin[†*] Yue Cao[*] Han Hu[*‡] Yixuan Wei[†]
Zheng Zhang Stephen Lin Baining Guo
Microsoft Research Asia
{v-zeliu1,v-yutlin,yuecao,hanhu,v-yixwe,zhez,stevelin,bainguo}@microsoft.com

김진성
Intelligence Engineering Lab, 고려대학교

Abstract

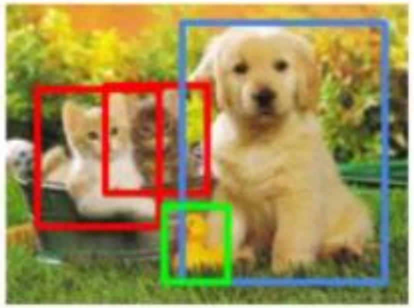
- This paper presents a new vision Transformer, called Swin Transformer, that capably serves as a general-purpose backbone for computer vision.
- Challenges in adapting Transformer from language to vision arise from differences between the two domains, such as large variations in the scale of visual entities and the high resolution of pixels in images compared to words in text.
- To address these differences, we propose a hierarchical Transformer whose representation is computed with Shifted windows.
- The shifted windowing scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows while also allowing for cross-window connection.
- This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to image size.
- These qualities of Swin Transformer make it compatible with a broad range of vision tasks, including image classification (87.3 top-1 accuracy on ImageNet-1K) and dense prediction tasks such as object detection (58.7 box AP and 51.1 mask AP on COCO test dev) and semantic segmentation (53.5 mIoU on ADE20K val).
- Its performance surpasses the previous state-of-the art by a large margin of +2.7 box AP and +2.6 mask AP on COCO, and +3.2 mIoU on ADE20K, demonstrating the potential of Transformer-based models as vision backbones.
- The hierarchical design and the shifted window approach also prove beneficial for all-MLP architectures.
- The code and models are publicly available at https://github.com/microsoft/Swin-Transformer..

Challenges in adapting Transformer to vision

Difference in language domain and visual domain

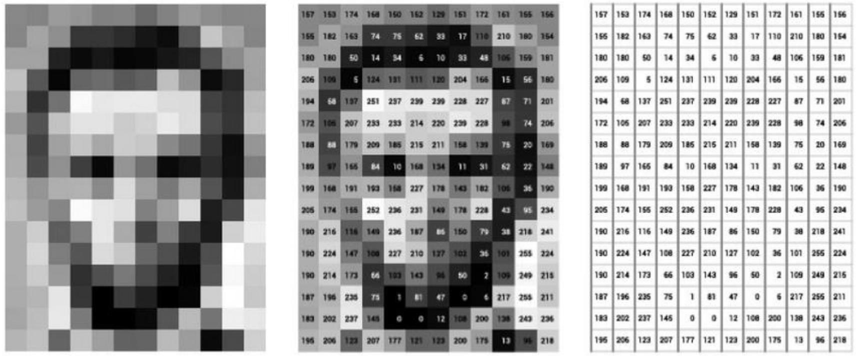
- One of these difference involves scale.
  - Visual elements can vary substantially in scale.

CAT



CAT, DOG, DUCK

- Another difference is the much higher resolution of pixels in images compared to words in passages of text.
  - Pixel level task would be intractable for Transformer on high resolution.



ViT(Vision Transformer)

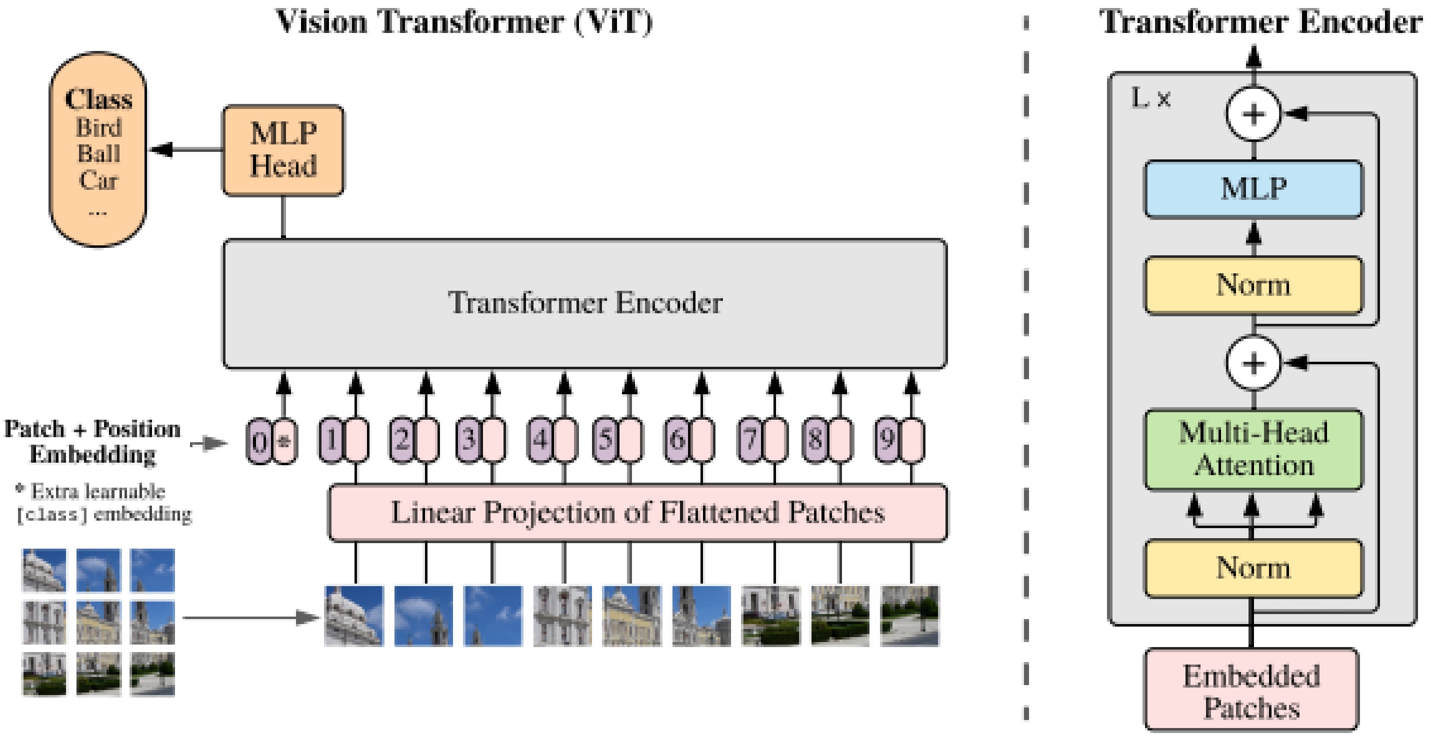
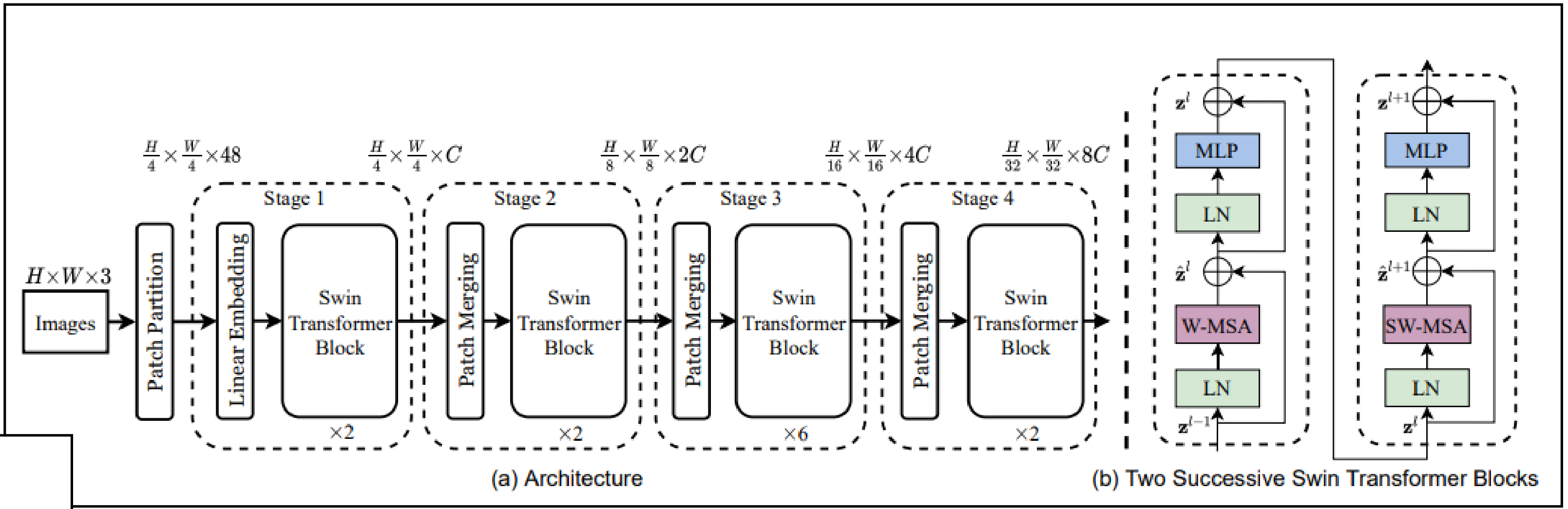


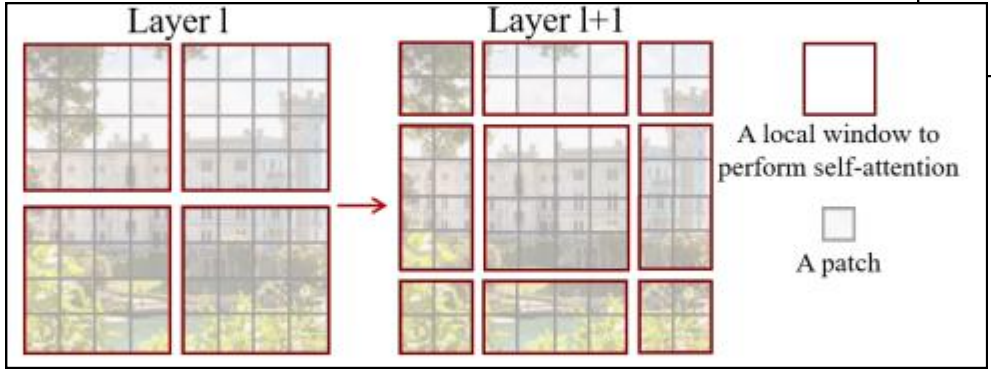
Figure 1: Model overview. We split an image into fixed-size patches, linearly embed each of them, add position embeddings, and feed the resulting sequence of vectors to a standard Transformer encoder. In order to perform classification, we use the standard approach of adding an extra learnable "classification token" to the sequence. The illustration of the Transformer encoder was inspired by Vaswani et al. (2017).

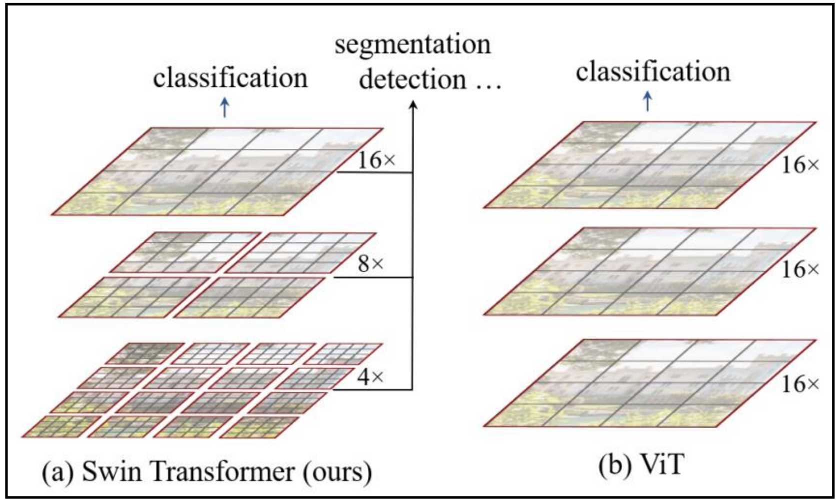
Swin Transformer.

1. Hierarchical Feature map
   - Patch Merging

2. Shifted window based Self-Attention
   - Window based self-attention
   - Shifted Window







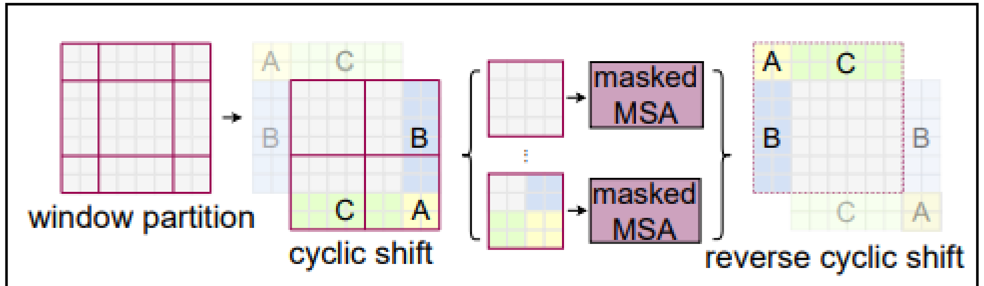
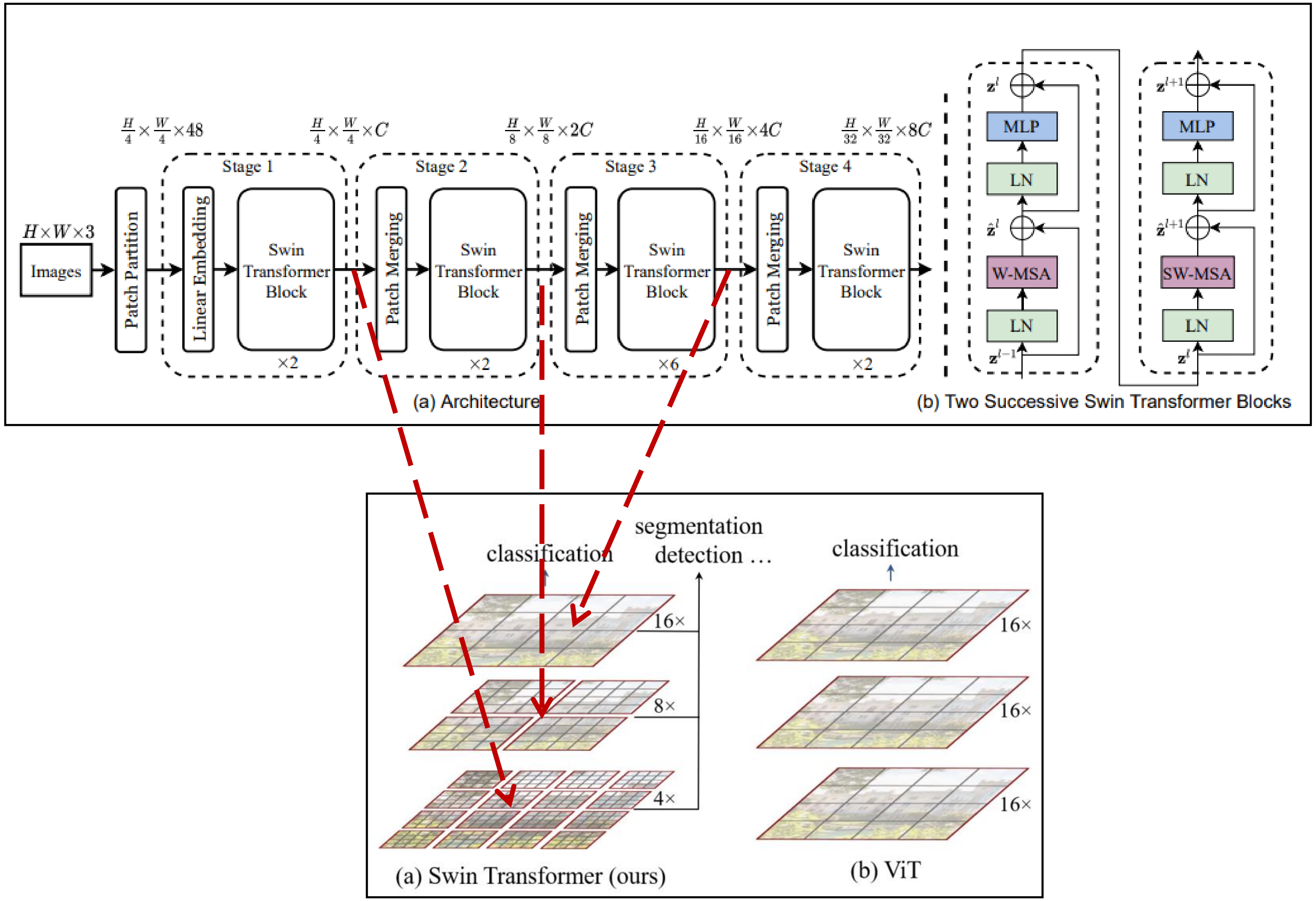


Figure 4. Illustration of an efficient batch computation approach for self-attention in shifted window partitioning.

Swin Transformer.

1. **Hierarchical Feature map**
   - Patch Merging

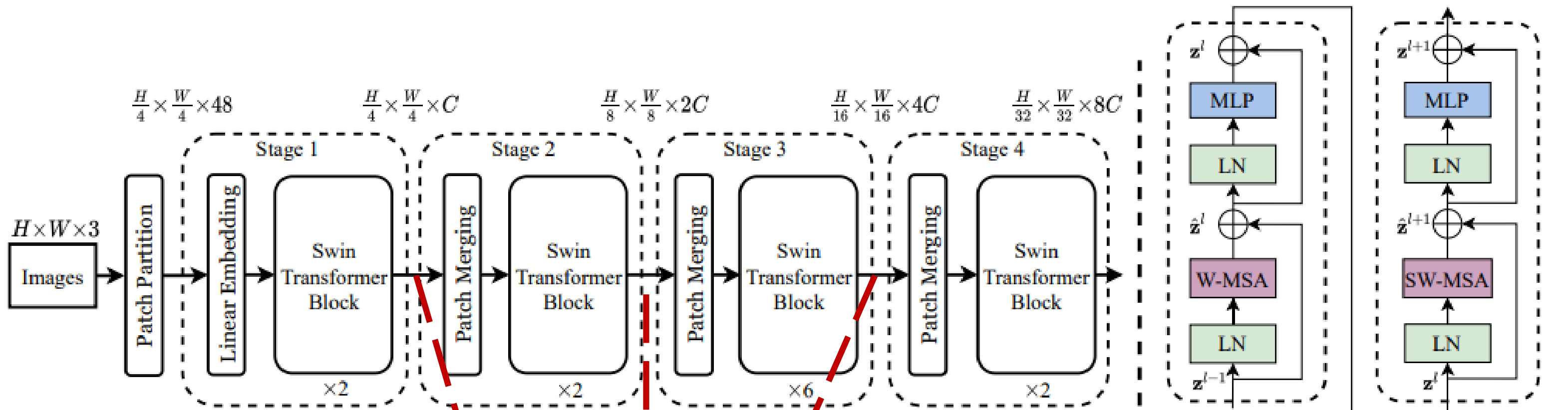
2. Shifted window based Self-Attention
   - Window based self-attention
   - Shifted Window



Swin Transformer.

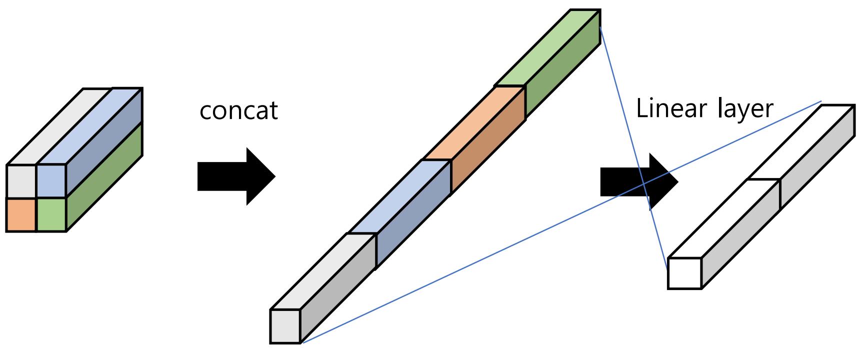
1. Hierarchical Feature map
   - **Patch Merging**

2. Shifted window based Self-Attention
   - Window based self-attention
   - Shifted Window

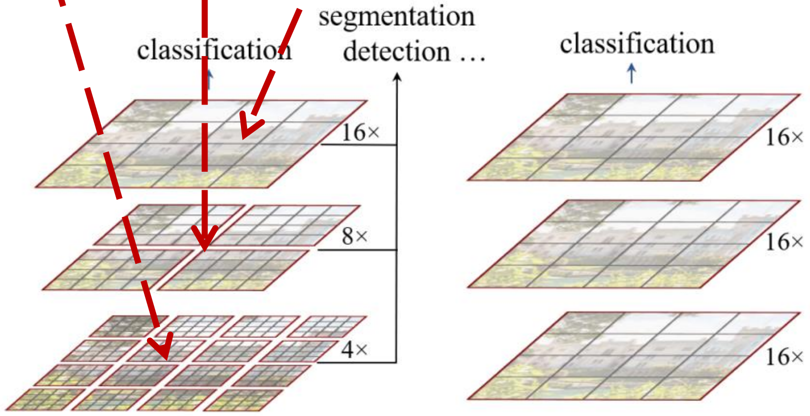


(a) Architecture

(b) Two Successive Swin Transformer Blocks



Patch Merging



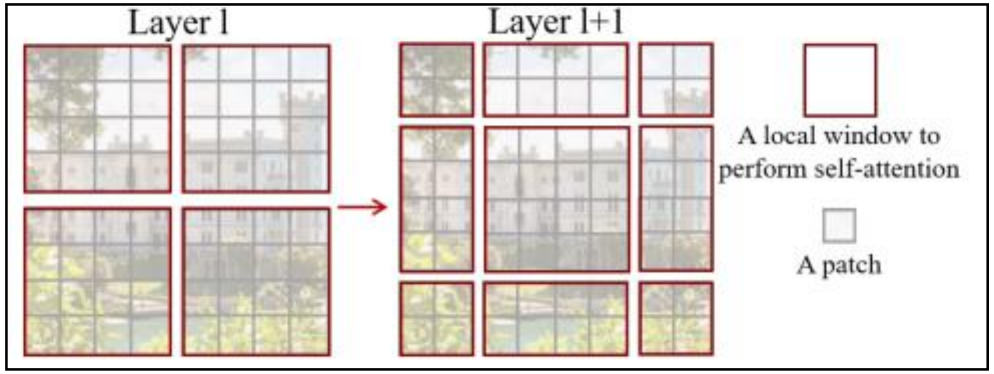
(a) Swin Transformer (ours)

(b) ViT

Swin Transformer.

1. Hierarchical Feature map
   - Patch Merging

2. Shifted window based Self-Attention
   - **Window based self-attention**
   - Shifted Window



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

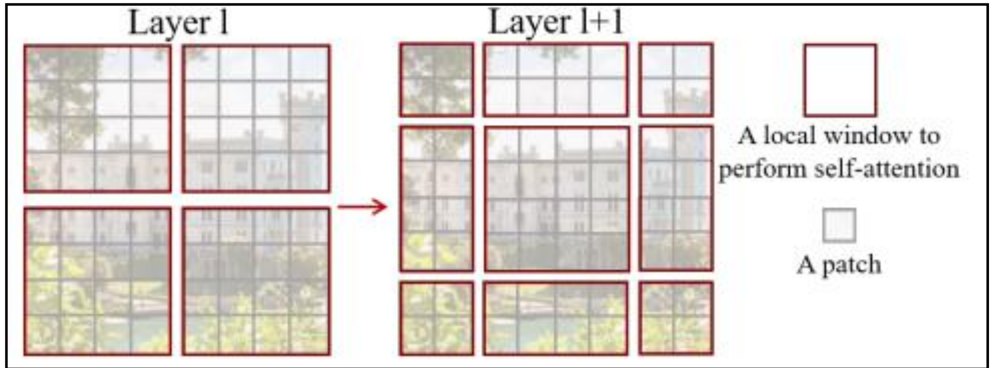
$$\Omega(\text{W-MSA}) = 4hwC^2 + 2M^2hwC, \quad (2)$$

where the former is quadratic to patch number hw, and the latter is linear when M is fixed (set to 7 by default). Global self-attention computation is generally unaffordable for a large hw, while the window based self-attention is scalable.

Swin Transformer.

1. Hierarchical Feature map
   - Patch Merging

2. Shifted window based Self-Attention
   - Window based self-attention
   - **Shifted Window**



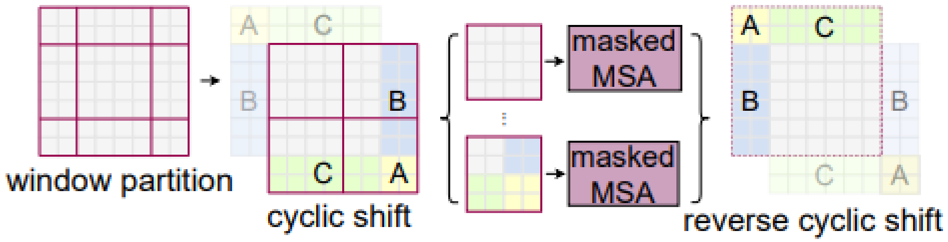


Figure 4. Illustration of an efficient batch computation approach for self-attention in shifted window partitioning.

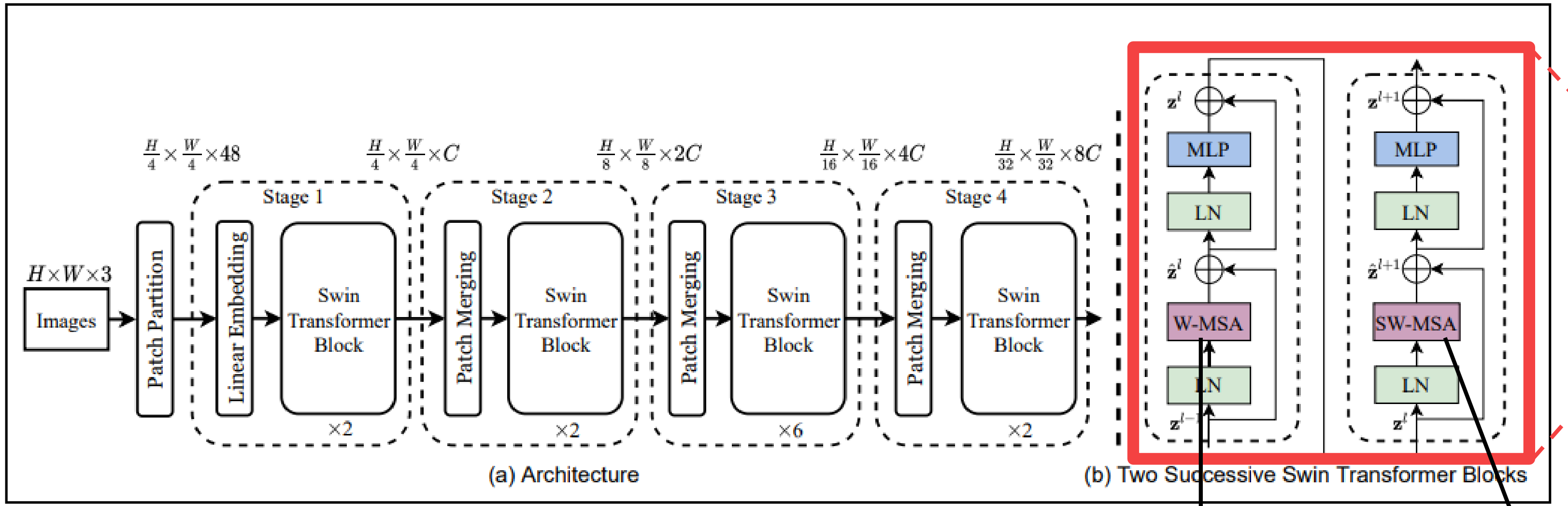
$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C, \quad (1)$$

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

where the former is quadratic to patch number hw, and the latter is linear when M is fixed (set to 7 by default). Global self-attention computation is generally unaffordable for a large hw, while the window based self-attention is scalable.

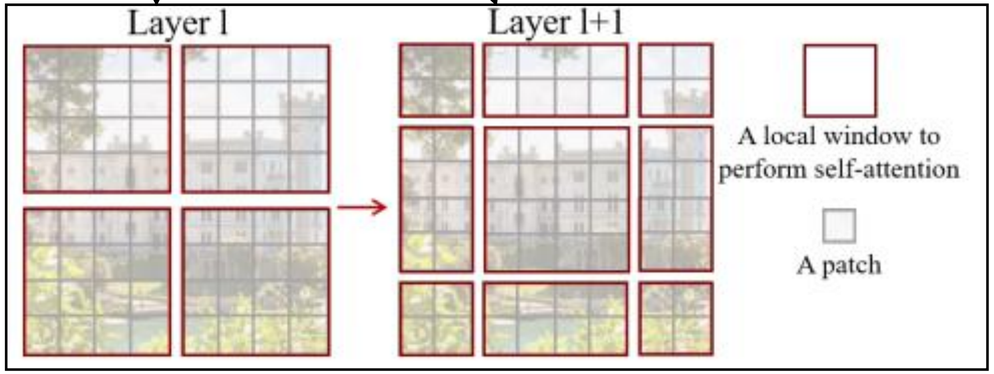
Swin Transformer.

1. Overall…



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

$$
\begin{aligned}
\hat{\mathbf{z}}^{l} &= \text{W-MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1}, \\
\mathbf{z}^{l} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l}\right)\right) + \hat{\mathbf{z}}^{l}, \\
\hat{\mathbf{z}}^{l+1} &= \text{SW-MSA}\left(\text{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l}, \\
\mathbf{z}^{l+1} &= \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1},
\end{aligned}
$$



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

QnA