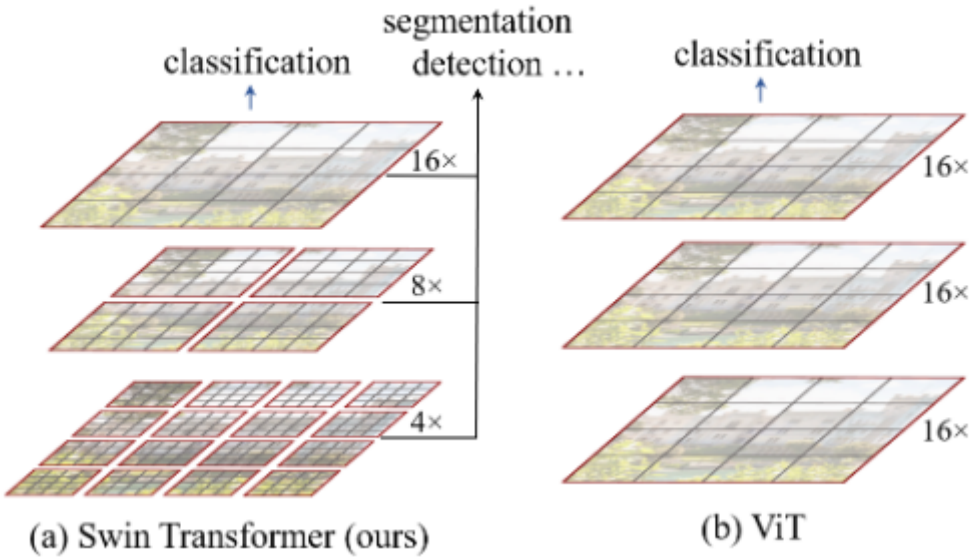
Swin Transformer

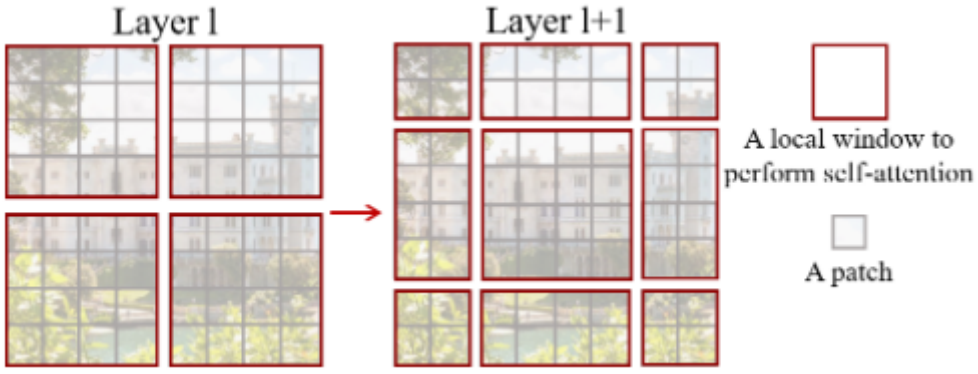
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows (2021)

4.8k

Hierarchical feature maps by merging image patches



Shifted window approach for computing self-attention



ViT: classification

Swin Transformer: general-purpose

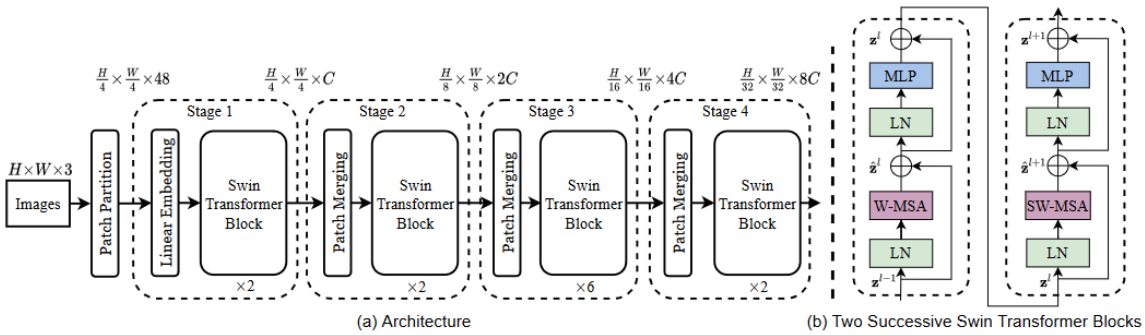


Figure 3. (a) The architecture of a Swin Transformer (Swin-T); (b) two successive Swin Transformer Blocks (notation presented with Eq. (3)). W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively.

Patch Merging

Similar to pooling in CNN

H * W * C → H/2 * W/2 * 4C → H/2 * W/2 * 2C

Shifted Window based Self-Attention

Self-attention in non-overlapped windows: quadratic complexity with respect to the number of tokens if global computation → linear complexity

Shifted window partitioning in successive blocks: window-based self-attention module lacks cross-window connections → regular window partitioning (first module) + shifted window partitioning (next module)

Efficient batch computation for shifted configuration: to avoid more windows in shifted window partitioning; cyclic shift + masked multi-head self-attention, reverse cyclic shift