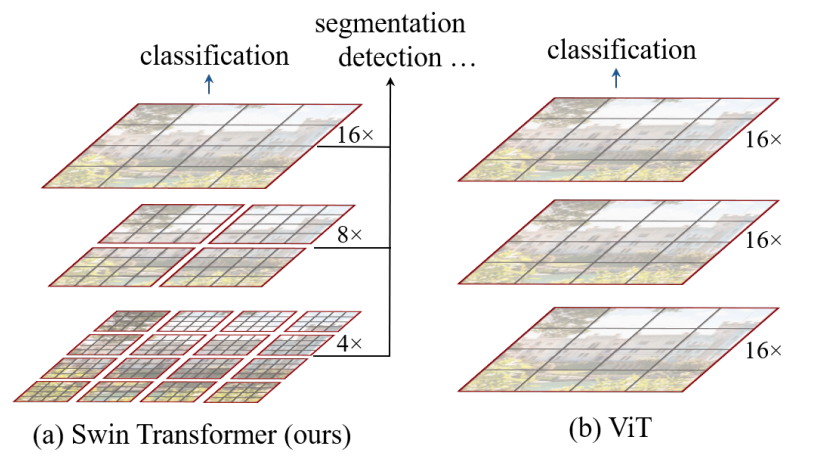
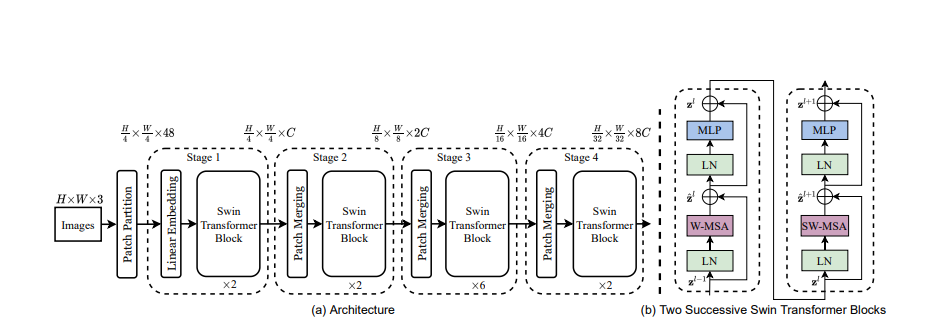
**Swin Transformer**

* ****The proposed Swin Transformer builds hierarchical feature maps by merging image patches (shown in gray) in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local window (shown in red). It can thus serve as a general-purpose backbone for both image classification and dense recognition tasks. (b) In contrast, previous vision Transformers produce feature maps of a single low resolution and have quadratic computation complexity to input image size due to computation of self-attention globally.
* 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.
* Swin Transformer constructs a hierarchical representation by starting from small-sized patches and gradually merging neighboring patches in deeper Transformer layers.
* A key design element of Swin Transformer is its shift of the window partition between consecutive self-attention layers.
* This strategy is also efficient in regards to real-world latency: all query patches within a window share the same key set1, which facilitates memory access in hardware. In contrast, earlier sliding window based self-attention approaches suffer from low latency on general hardware due to different key sets for different query pixels.
* An overview of the Swin Transformer architecture is presented in Figure, which illustrates the tiny version (SwinT). It first splits an input RGB image into non-overlapping patches by a patch splitting module, like ViT. Each patch is treated as a “token” and its feature is set as a concatenation of the raw pixel RGB values. 
* Several Transformer blocks with modified self-attention computation (Swin Transformer blocks) are applied on these patch tokens. The Transformer blocks maintain the number of tokens (H 4 × W 4), and together with the linear embedding are referred to as “Stage 1”.
* To produce a hierarchical representation, the number of tokens is reduced by patch merging layers as the network gets deeper. The first patch merging layer concatenates the features of each group of 2 × 2 neighboring patches and applies a linear layer on the 4C-dimensional concatenated features. This reduces the number of tokens by a multiple of 2×2 = 4 (2× down sampling of resolution), and the output dimension is set to 2C. Swin Transformer blocks are applied afterwards for feature transformation, with the resolution kept at H 8 × W 8. This first block of patch merging and feature transformation is denoted as “Stage 2”.
* Swin Transformer is built by replacing the standard multi-head self-attention (MSA) module in a Transformer block by a module based on shifted windows with other layers kept the same. Swin Transformer block consists of a shifted window based MSA module, followed by a 2-layer MLP with GELU nonlinearity in between. A Layer Norm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module.
* The standard Transformer architecture and its adaptation for image classification both conduct global self-attention, where the relationships between a token and all other tokens are computed. The global computation leads to quadratic complexity with respect to the number of tokens, making it unsuitable for many vision problems requiring an immense set of tokens for dense prediction or to represent a high-resolution image.