An overview of the Swin Transformer architecture is presented as following. It first splits an input RGB image into non-overlapping patches by a patch splitting module, similar to ViT. Each patch is treated as a “token” and its feature is set as a concatenation of the raw pixel RGB values. In the Swin, the patch size is 4 × 4 and thus the feature dimension of each patch is 4 × 4 × 3 = 48. A linear embedding layer is applied on this raw-valued feature to project it to an arbitrary dimension (denoted as C).

Several Swin Transformer blocks are applied on these patch tokens. These blocks maintain the number of tokens ( H/4 × W/4 ), and together with the linear embedding are referred to as “Stage 1”. Then, this model repeats the procedure three times, as “Stage 2”, “Stage 3” and “Stage 4”, with output resolutions of H/8 × W/8, H/16 × W/16 and H/32 × W/32, respectively. These stages jointly produce a hierarchical representation.

**Swin Transformer block**

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. As illustrated in Figure 1(b), a Swin Transformer block consists of a shifted window based MSA module, followed by a 2-layer MLP with GELU nonlinearity in between. A LayerNorm (LN) layer is applied before each MSA module and each MLP, and a residual connection is applied after each module.

**Shifted window**

As illustrated in Figure 2. The shifted windows bridge the windows of the preceding layer, providing connections among them that significantly enhance modeling power. In layer l, a regular window partitioning scheme is adopted, and self-attention is computed within each window. In the next layer l + 1, the window partitioning is shifted, resulting in new windows. The self-attention computation in the new windows crosses the boundaries of the previous windows in layer l, providing connections among them.

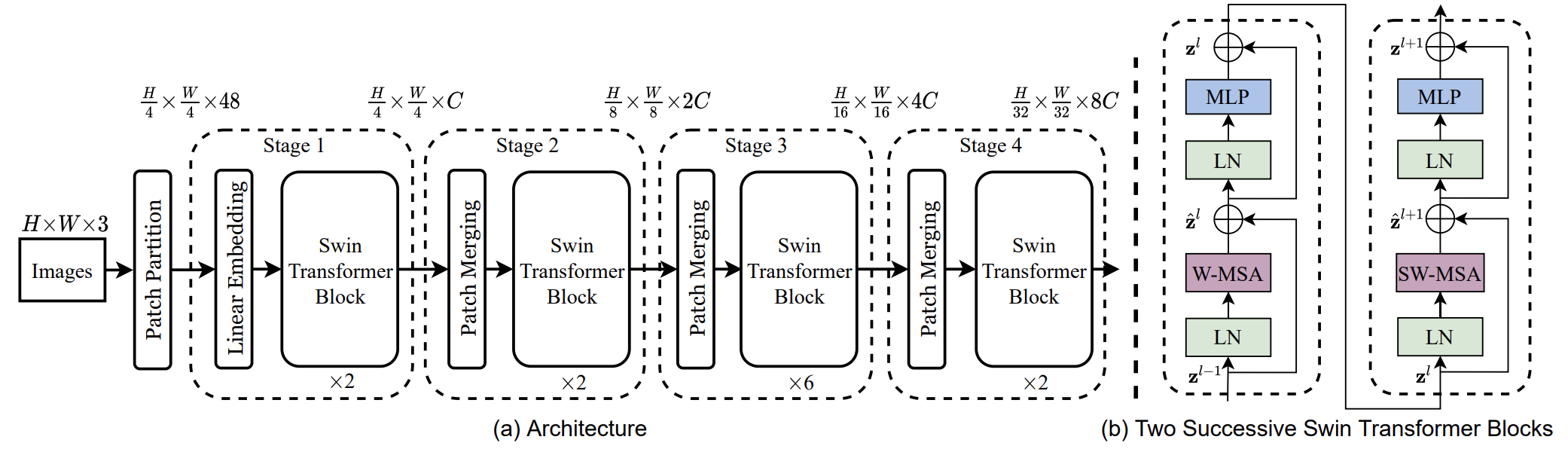


Figure1: (a) The architecture of a Swin Transformer; (b) two successive Swin Transformer Blocks. W-MSA and SW-MSA are multi-head self attention modules with regular and shifted windowing configurations, respectively

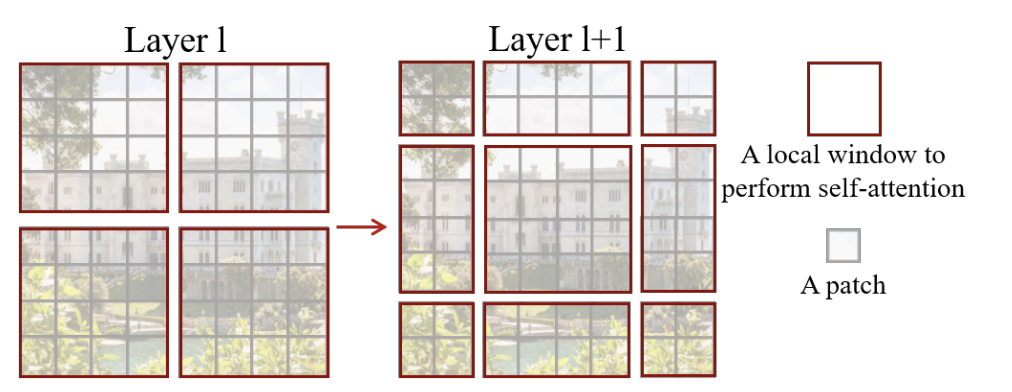


Figure2: An illustration of the shifted window approach for computing self-attention in the proposed Swin Transformer architecture.