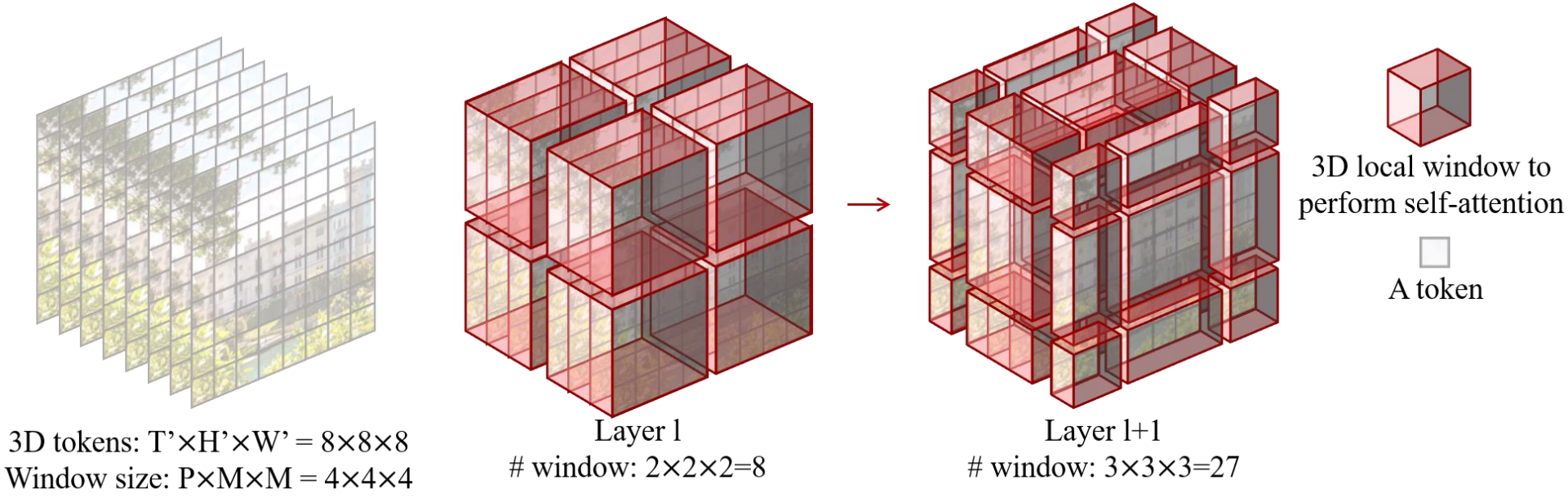
Video classification with video swin transformer

3/11

Shawson Hsiao

Overall Structure



- Videos is a sequence of images in temporal dimension
- Still have hierarchical design with patch merging and self-attention within a window but extends the scope from the spatial domain to the spatiotemporal domain.
- Each patch is a 3D token

Advantages

- **Spatiotemporal locality**:
  pixels that are closer to each other in the spatiotemporal distance are more likely to be correlated, no need full spatiotemporal self-attention

- Still preserve locality, hierarchy and translation invariance

Parameters

- Patch size : (4,4,4)
- embed_dim C: 96
- Window_size: (2,7,7)

```
def window_partition(x, window_size):
    """
    Args:
        x: (B, D, H, W, C)
        window_size (tuple[int]): window size

    Returns:
        windows: (B*num_windows, window_size*window_size, C)
    """
```