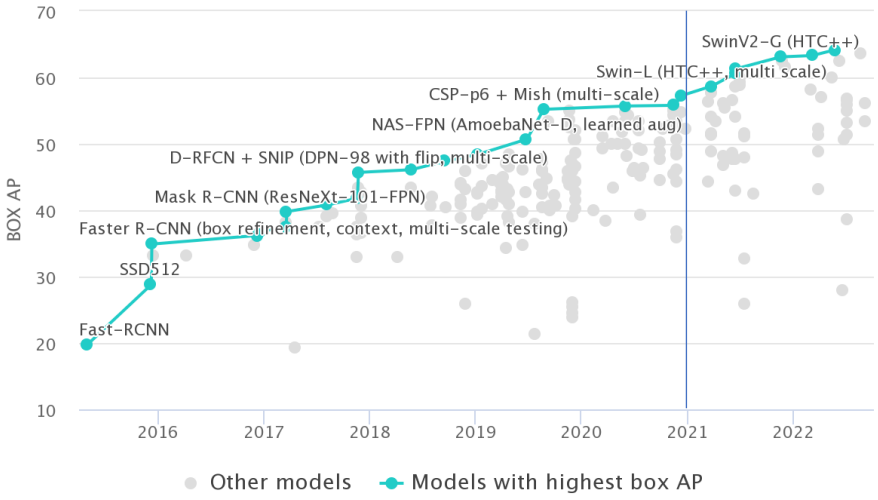
THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

Swin Transformer
Hierarchical Vision Transformer using Shifted Windows

Xiangqiao MENG 22041201r

Excellent Performance in Object Detection



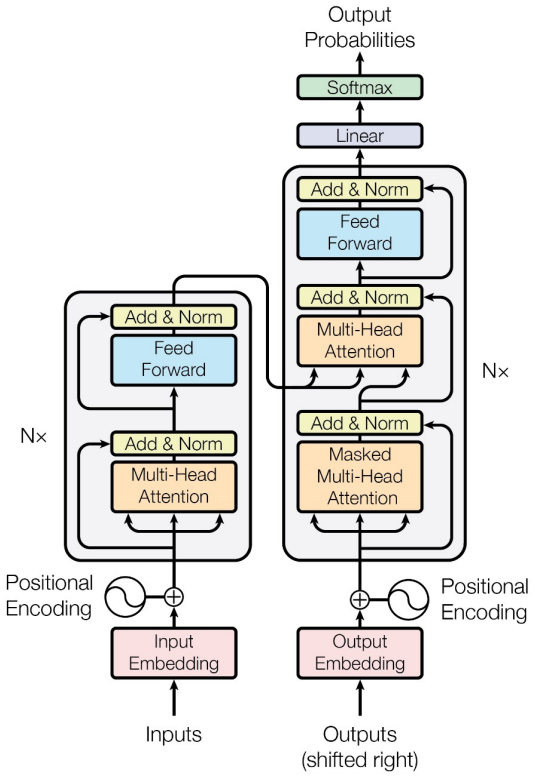
Leaderboard on COCO test-dev

Rank	Model	Box AP	Year
1	FD-SwinV2-G	64.2	2022
2	BEiT-3	63.7	2022
3	DINO	63.3	2022
4	SwinV2-G	63.1	2021
5	Florence-CoSwin-H	62.4	2021
6	GLIPv2	62.4	2022
7	GLIP	61.5	2022
8	Soft Teacher + Swin-L	61.3	2021
9	DyHead	60.6	2021
10	ViT-Adapter-L	60.1	2022

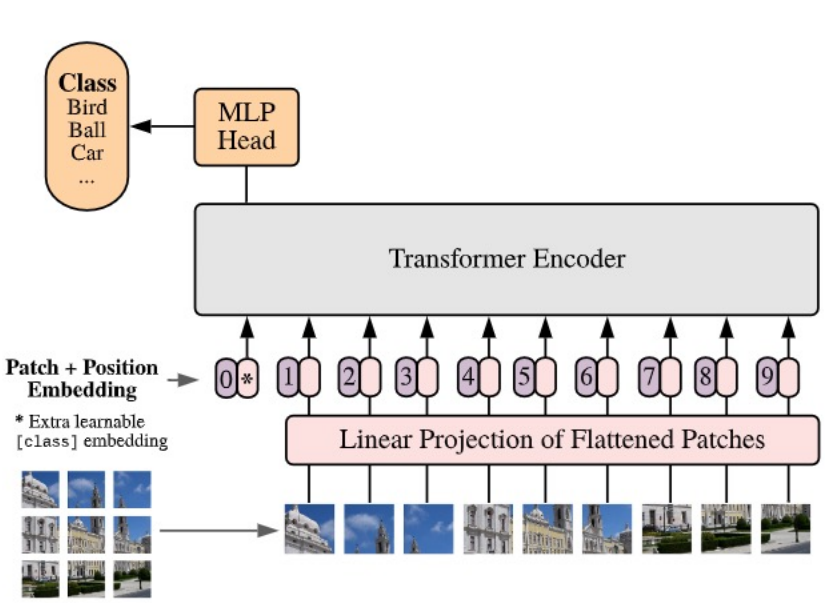
Top 10 Model on COCO test-dev

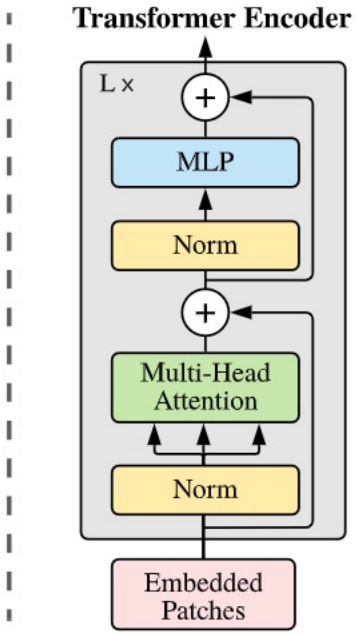
Image and Table Data Source: https://paperswithcode.com/sota/object-detection-on-coco

Transformer[1] and ViT[2]



Transformer[1]

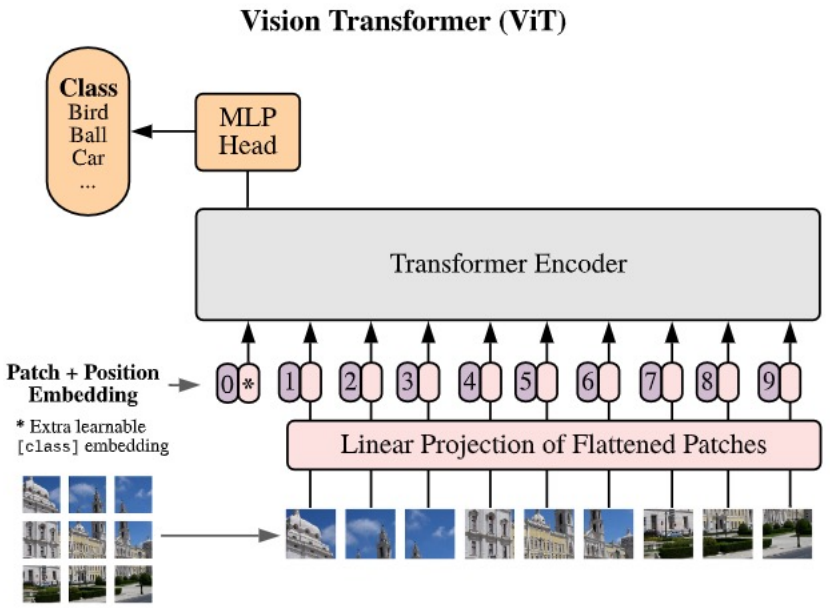




Vision Transformer (ViT)[2]

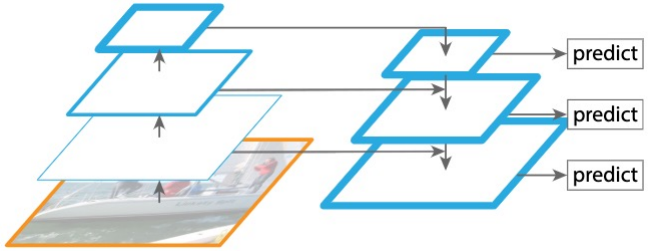
Vision Transformer – ViT[2]

- Single-Scale Features
- Low Resolution
- Quadratic Complexity

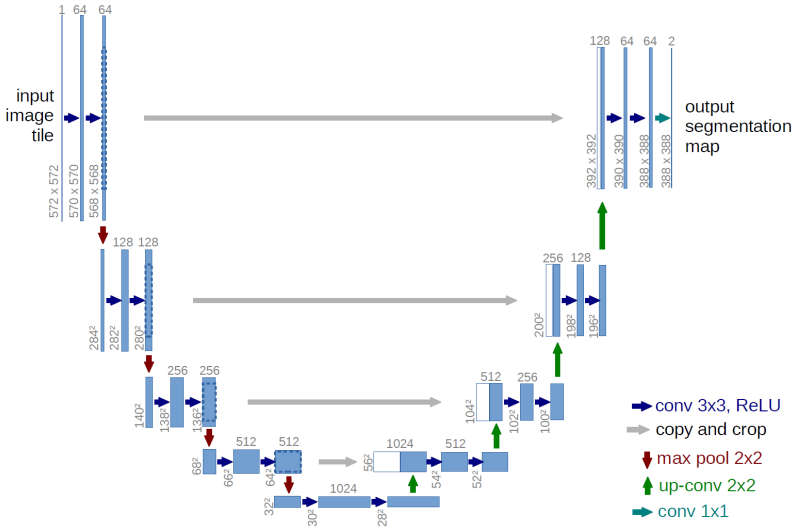


Multi-Scale Feature in CNN

- Feature Pyramid Networks (FPN)[4]

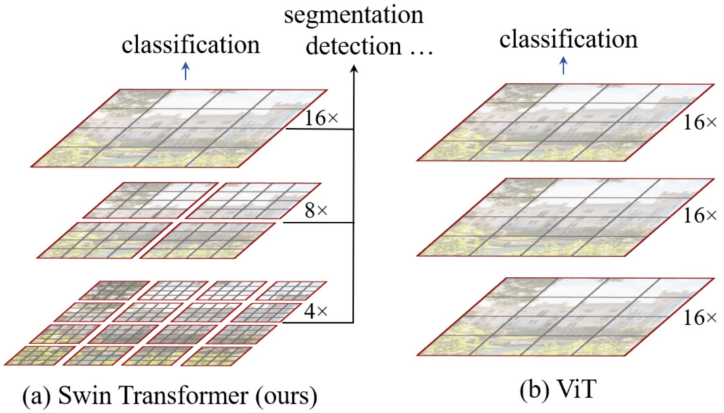


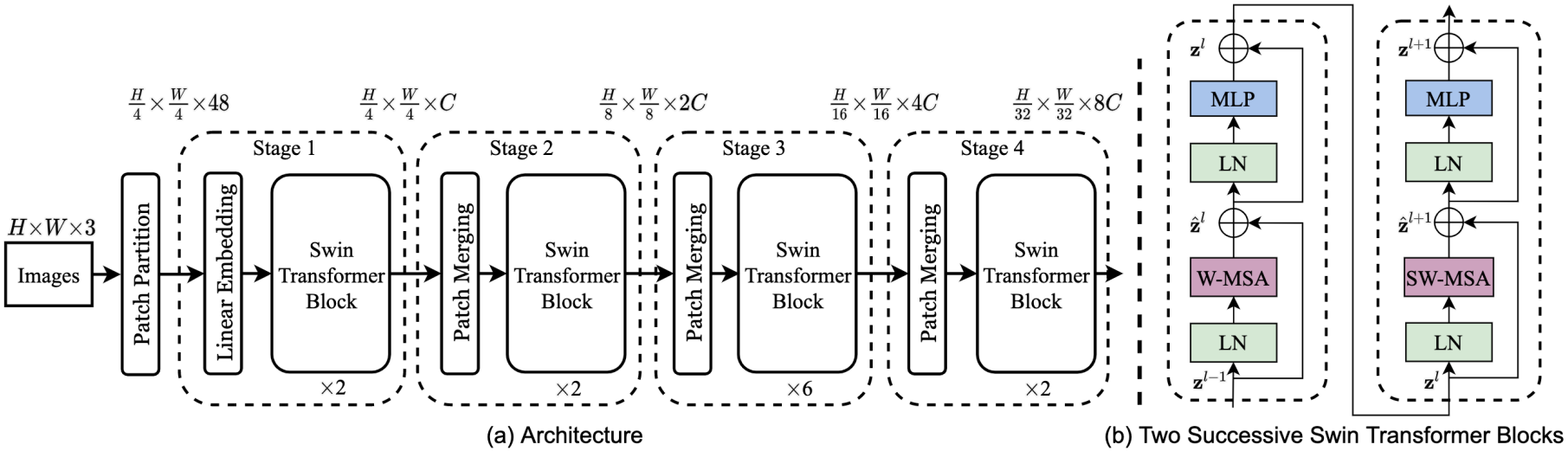
- U-Net



Swin Transformer

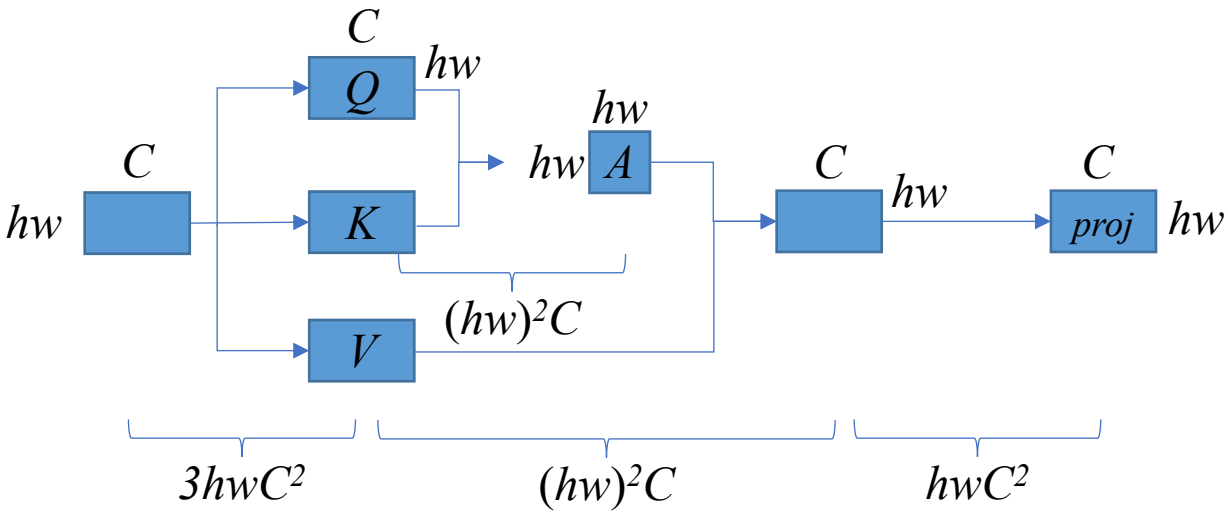
- Backbone





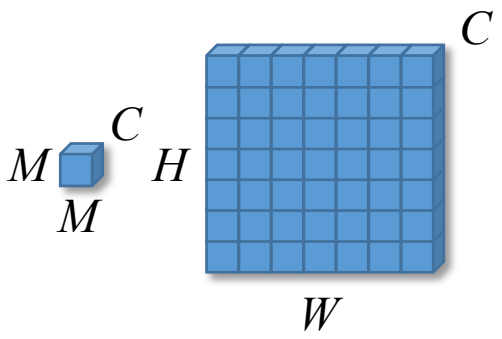
Swin Transformer

- Window based Self-Attention
  - Computational Complexity of MSA
    - $\Omega(MSA) = 4hwC^2 + 2(hw)^2C$



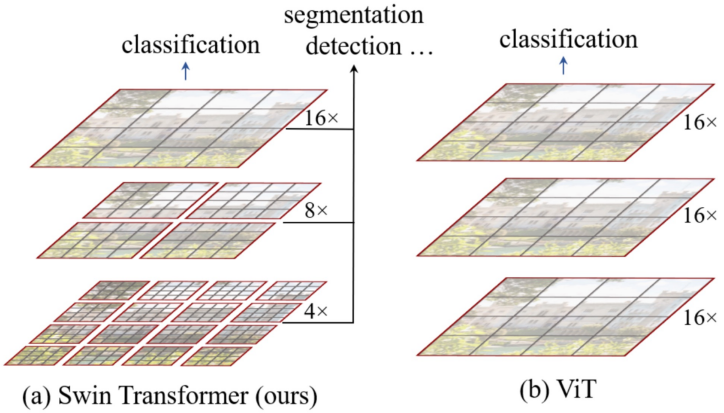
Swin Transformer

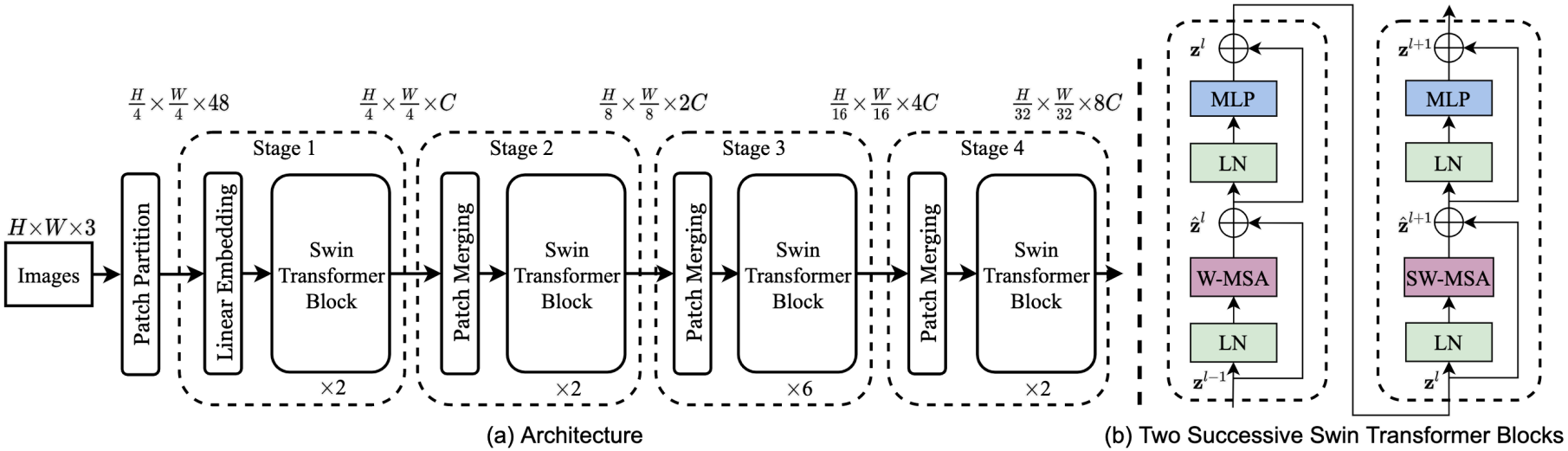
- Window based Self-Attention
  - Computational Complexity of MSA
    - $\Omega(MSA) = 4hwC^2 + 2(hw)^2C$
  - Self-Attention in each window (size = M x M)
    - Computational Complexity in 1 window
      - $\Omega(1\_window) = 4M^2C^2 + 2M^4C$
    - Total Computational Complexity
      - $\Omega(W\_MSA) = 4hwC^2 + 2M^2hwC$



Swin Transformer

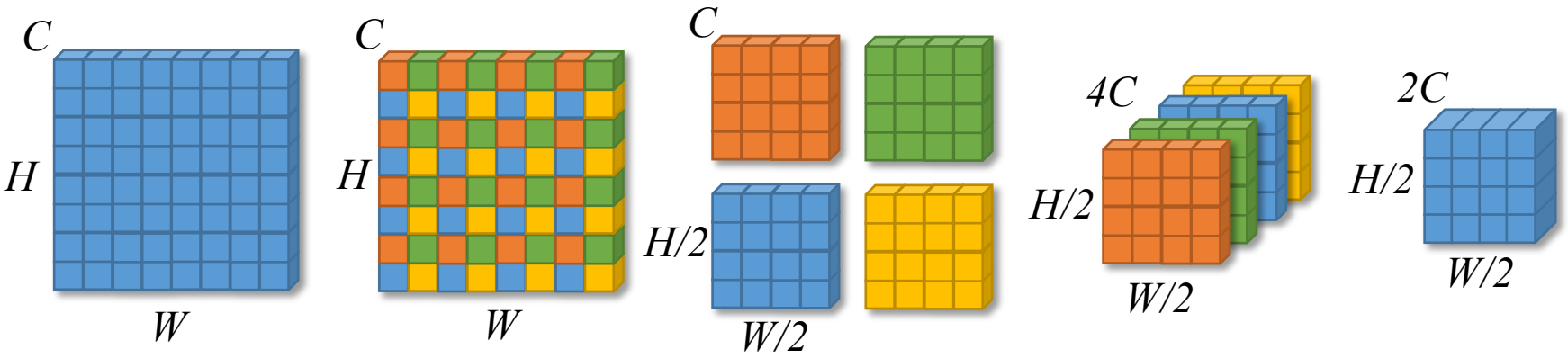
- Backbone





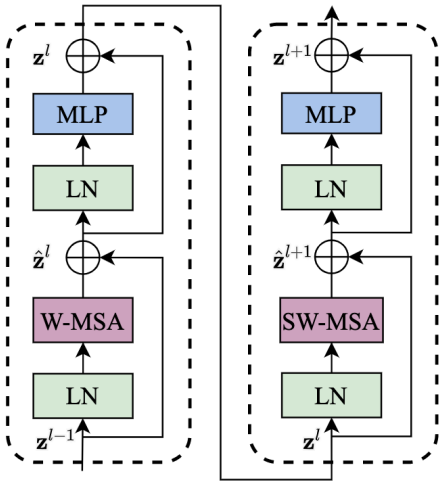
Swin Transformer

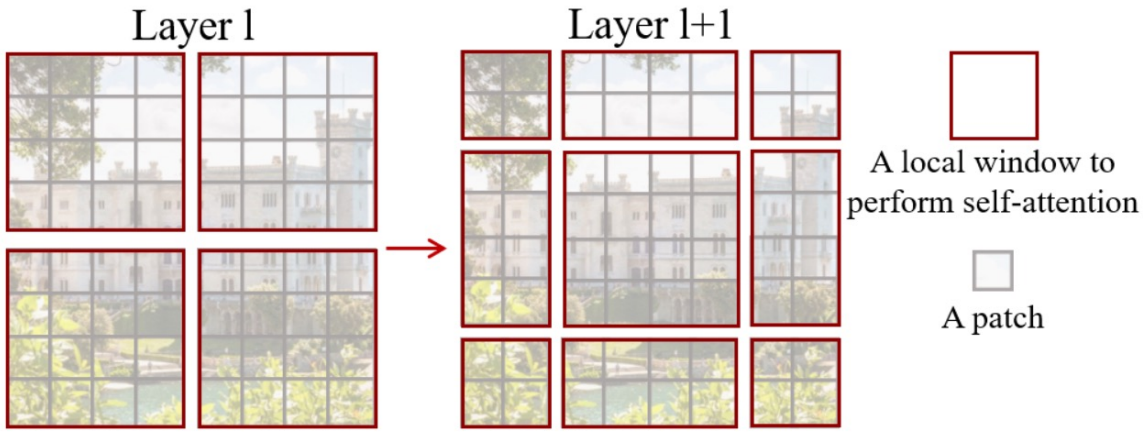
- Patch Merging



Swin Transformer

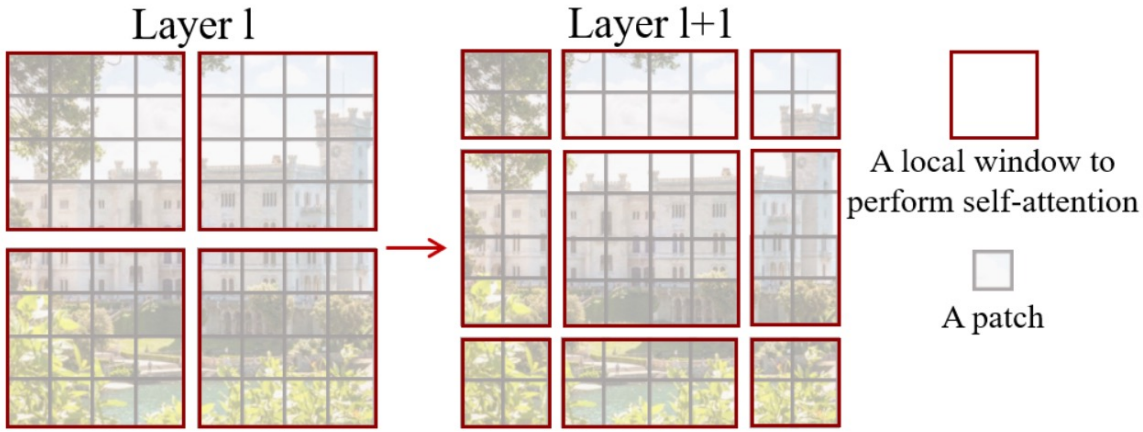
- Shifted Window
  - Introduce cross-window connections while maintaining the efficient computation of non-overlapping windows





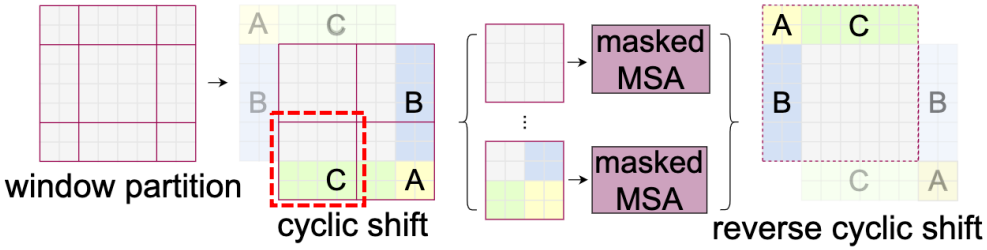
Swin Transformer

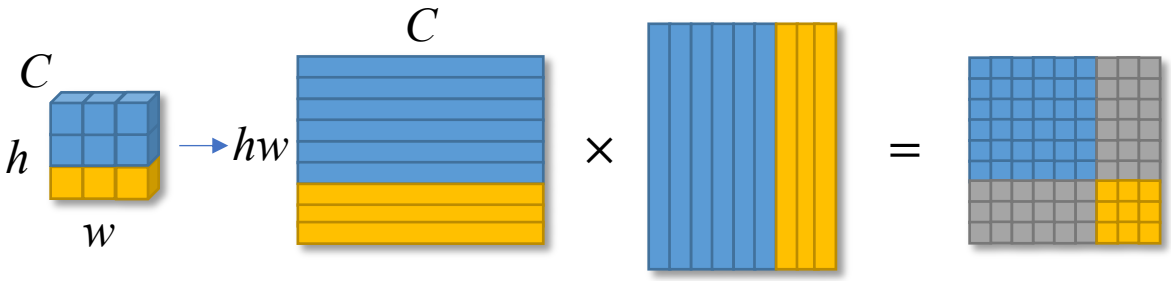
- Shifted Window
  - The number of windows changed.
  - The size of each window is different.



Swin Transformer

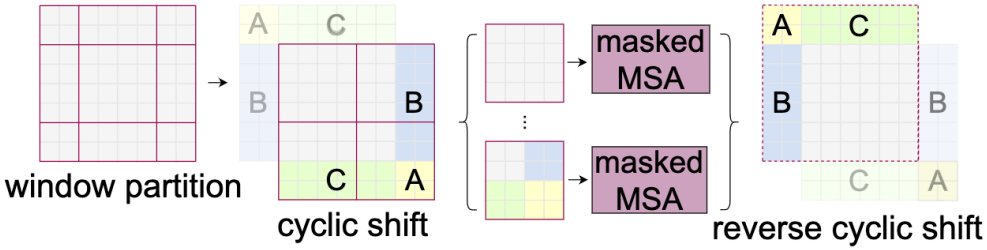
- Shift Window – Masked MSA





Swin Transformer

- Shift Window – Masked MSA



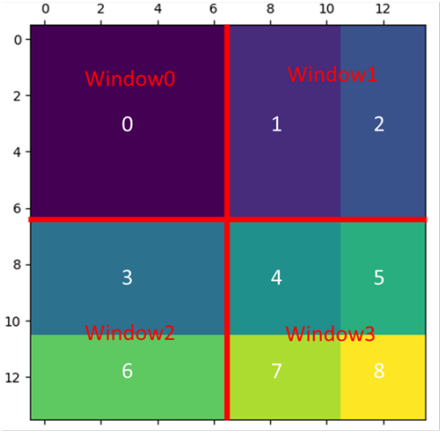
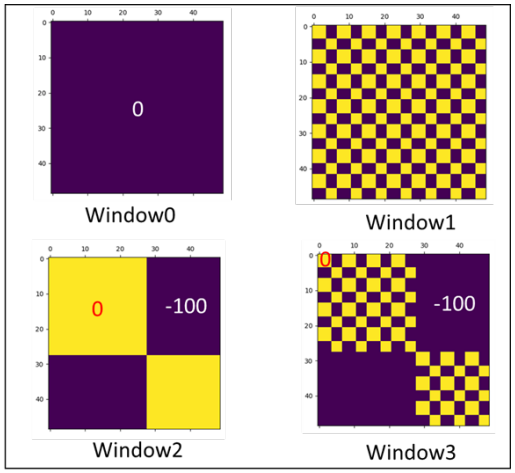


Image Mask
(14x14, window 7x7, shift 3)



Attn Mask

Reference

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[3] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[4] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

Thank you!