Supplementary Material for PanoSwin: a Pano-style Swin Transformer for Panorama Understanding

Zhixin Ling Zhen Xing Xiangdong Zhou Manliang Cao Guichun Zhou
School of Computer Science, Fudan University
{20212010005,zxing20,xdzhou,17110240029,19110240014}@fudan.edu.cn

1. Architecture Details

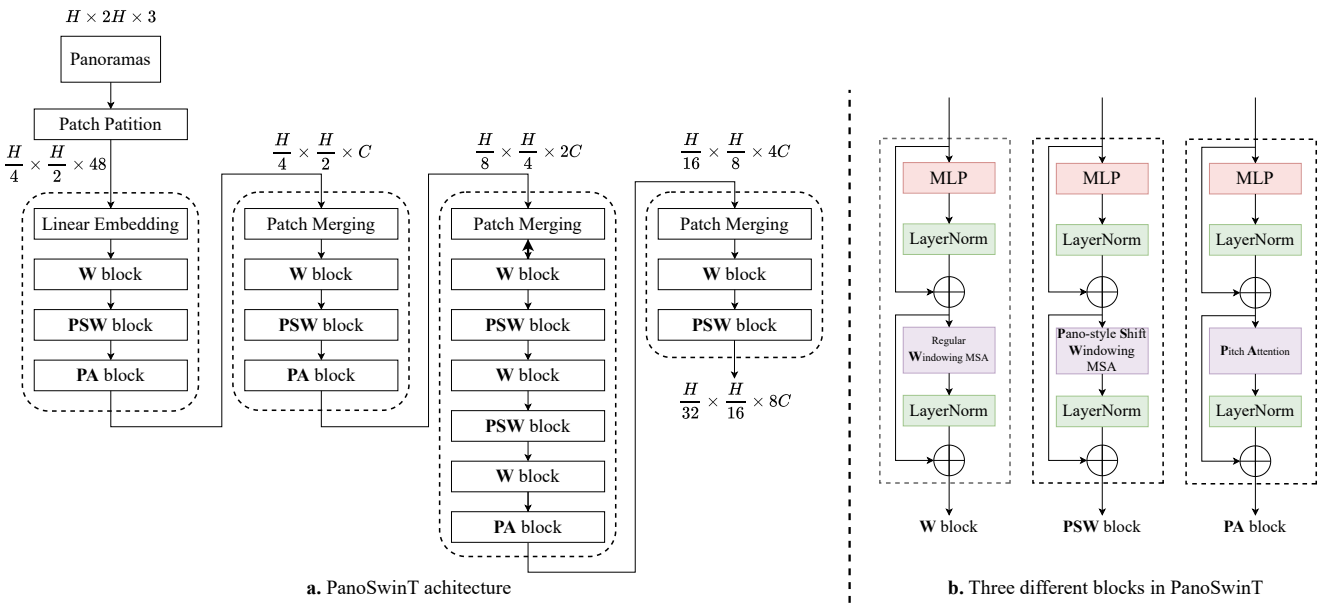
Tab.3 in the main submission briefly shows architectures of PanoSwin. We further explain the network architecture in Fig. 1. In the figure, Linear Embdding expands the feature dimension to C; C is the pre-defined "embedding dim" of Tab.3 in the main submission. Patch Merging [1] expands the feature channel and reduces the feature size.

Specifically, take PanoSwinT for an example, as shown in Fig. 1-**b**, there are three different blocks in PanoSwinT: the **W** block, the **PSW** block, and the **PA** block. There only differ in their multi-head self attention module: regular windowing attention for the **W** block [1], our proposed pano-style shift windowing attention for the **PSW** block, and our pitch attention for the **PA** block.

Similar to SwinT [1], there are mainly four stages in PanoSwinT, as shown in the rounded rectangles in Fig. 1-**a**. The Linear Embedding/Patch Merging module in each stage will expand feature dimensions. Besides, the Patch Merging module also reduce both feature width and height by half. In the PanoSwinT architecture, the **W** block and the **PSW** block are always composed together to enable inter-window interactions. The **PA** block is inserted to the last within the first three stages to remove spatial distortion.

References

[1] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 1



a. PanoSwinT achitecture

b. Three different blocks in PanoSwinT

Figure 1. Illustration of PanoSwin architecture.