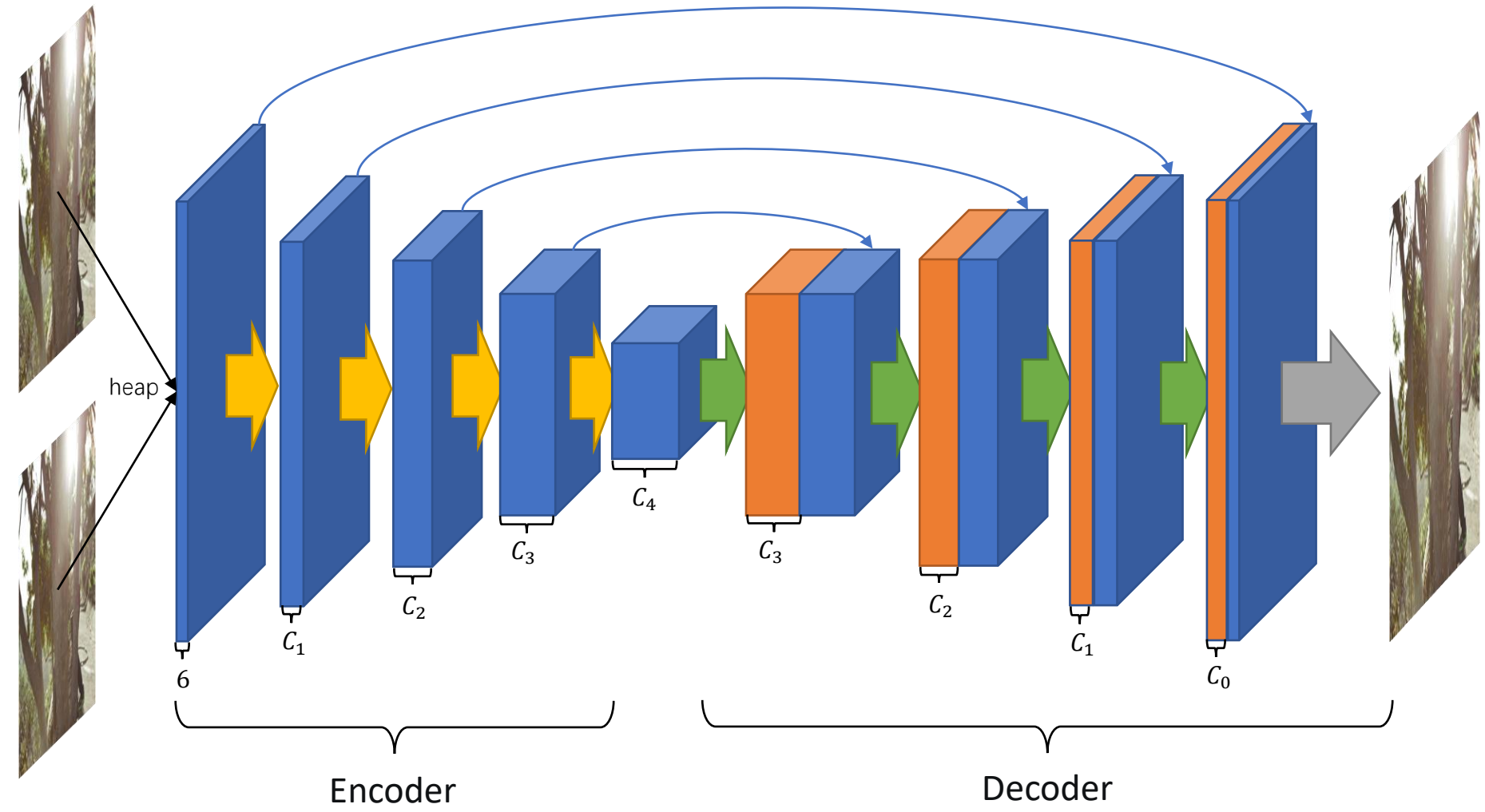
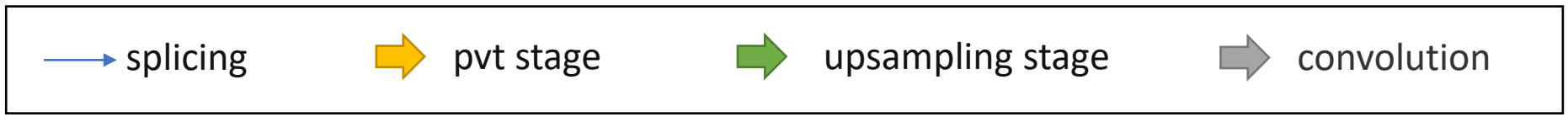
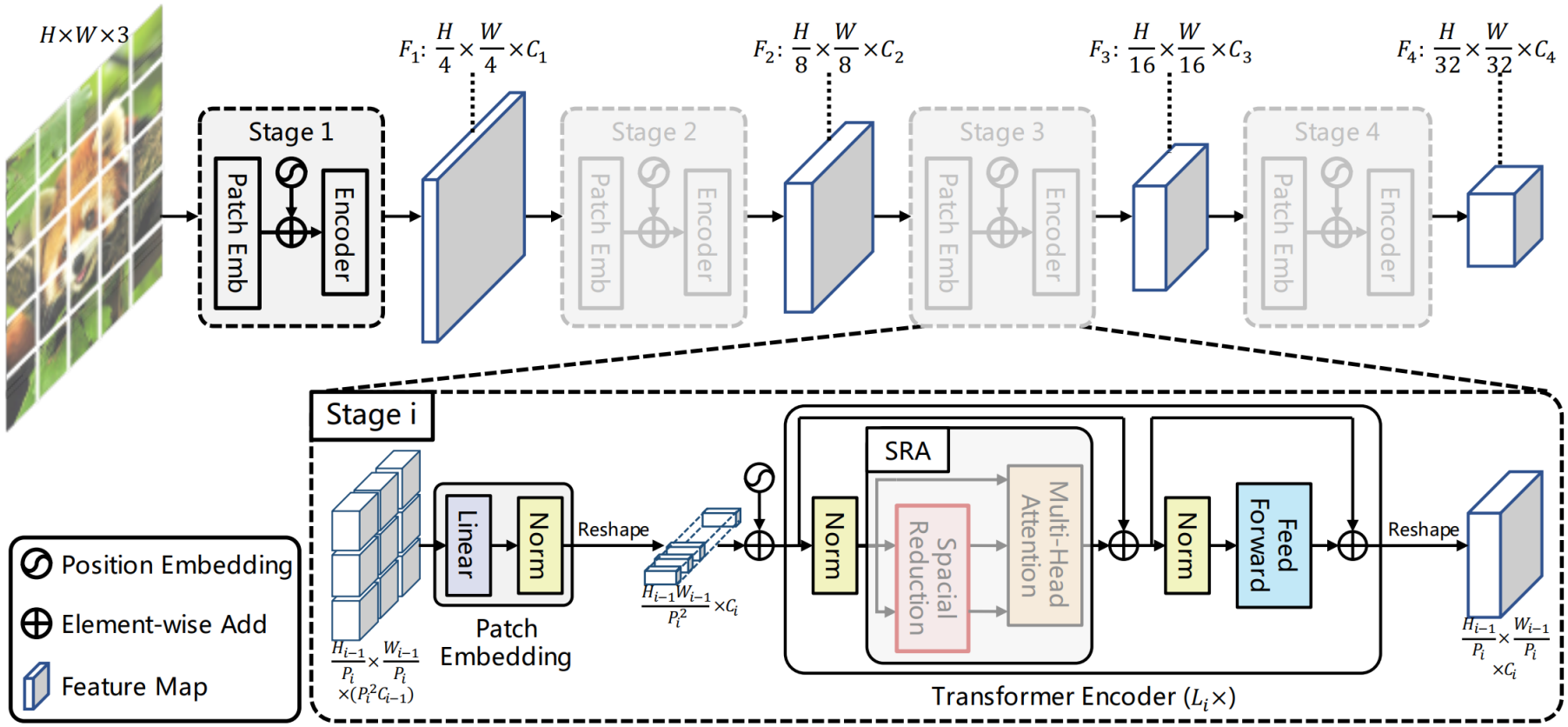
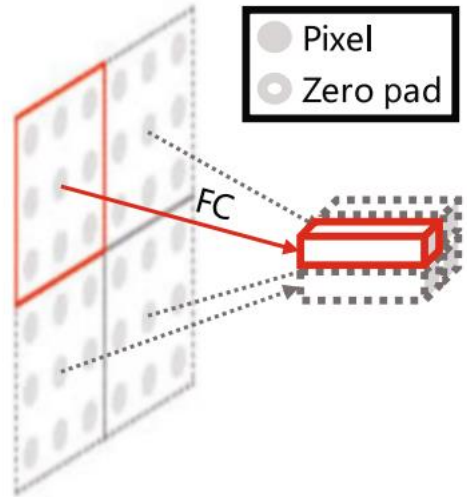




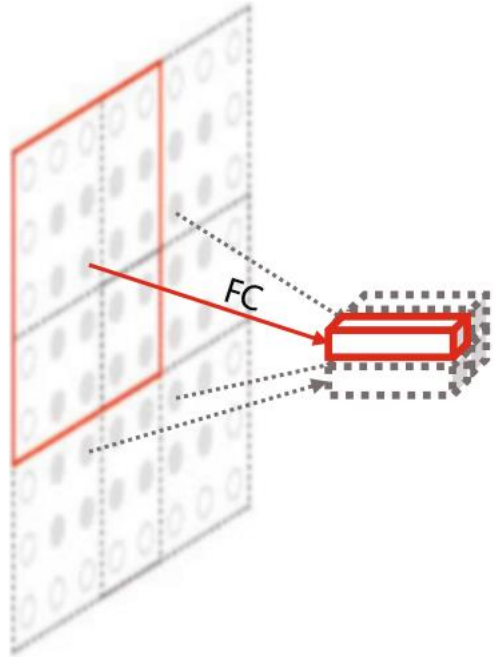
Noting:
The diagram of the upsampling stage in this figure is not entirely accurate, it is for illustrative purposes only. Please refer to the specific diagram for accuracy.

Encoder (picture from the original paper of pvt)



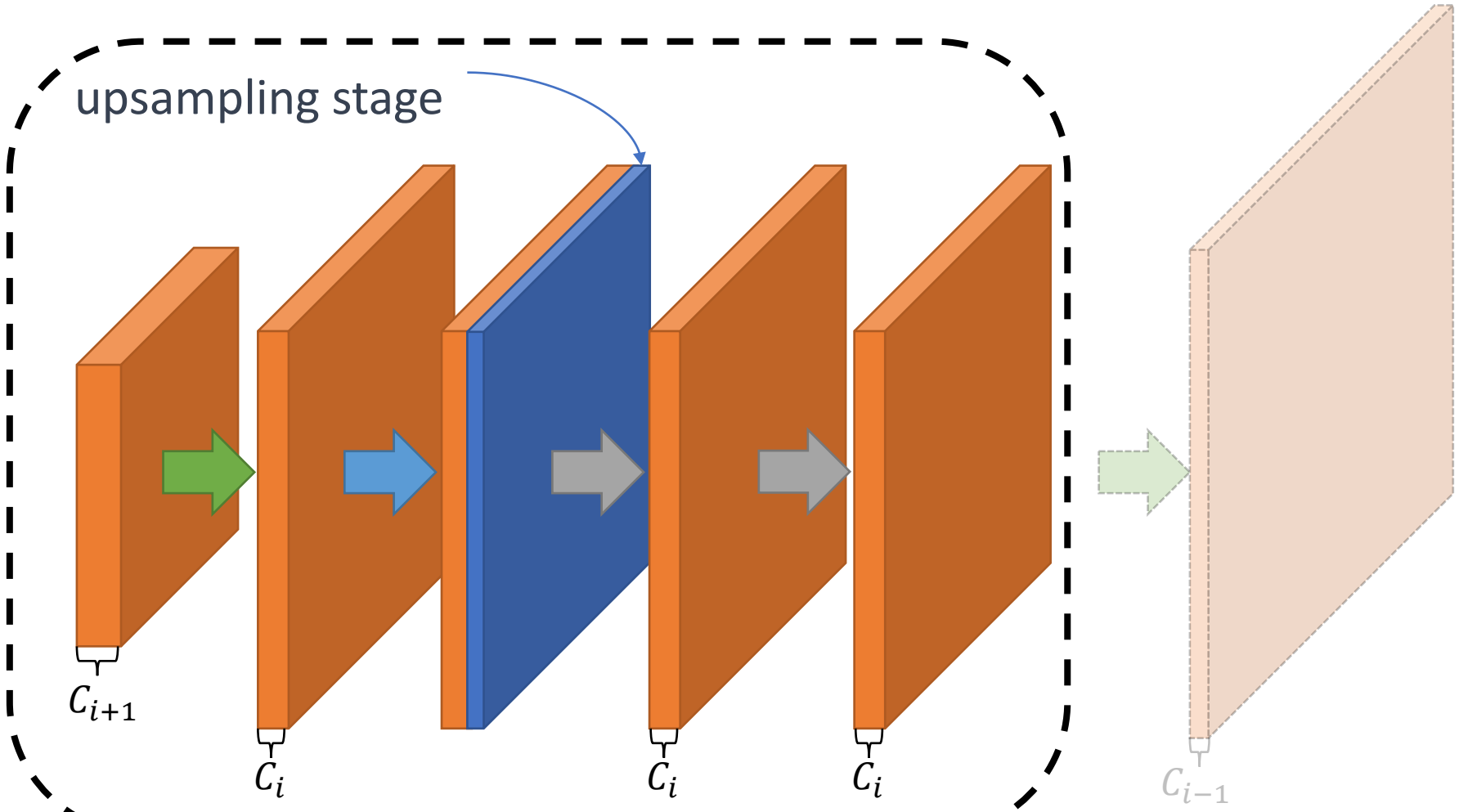


PVT v1: Original patch embedding



PVT v2: Overlapping patch embedding

Compared with the pvt structure above, my model used the overlapping patch embedding (picture on the right) and cancelled position codes. I also recorded every output of the stages to conduct splicing.



The channels of the blue part is not always equals to C_i. In my code, it can be 6 in the last upsampling stage.

deconvolution convolution splicing