Encoder

Decoder

$\frac{H}{4} \times \frac{W}{4} \times 4C$

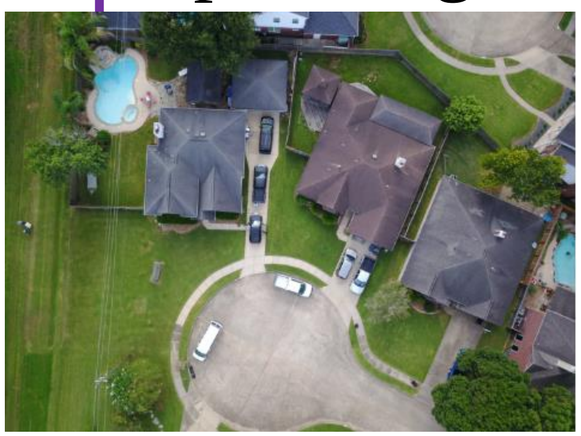
$\frac{H}{4} \times \frac{W}{4} \times Ncl_s$

MLP Layer

MLP

Predicted mask

Input image



$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$

MLP

$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C$

Up sample

$\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C$

Patch partitioning

Linear embedding

SwinT block

Stages: S_i where i ϵ {1, 2, 3,4},

Feature size at S_i: $\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i$