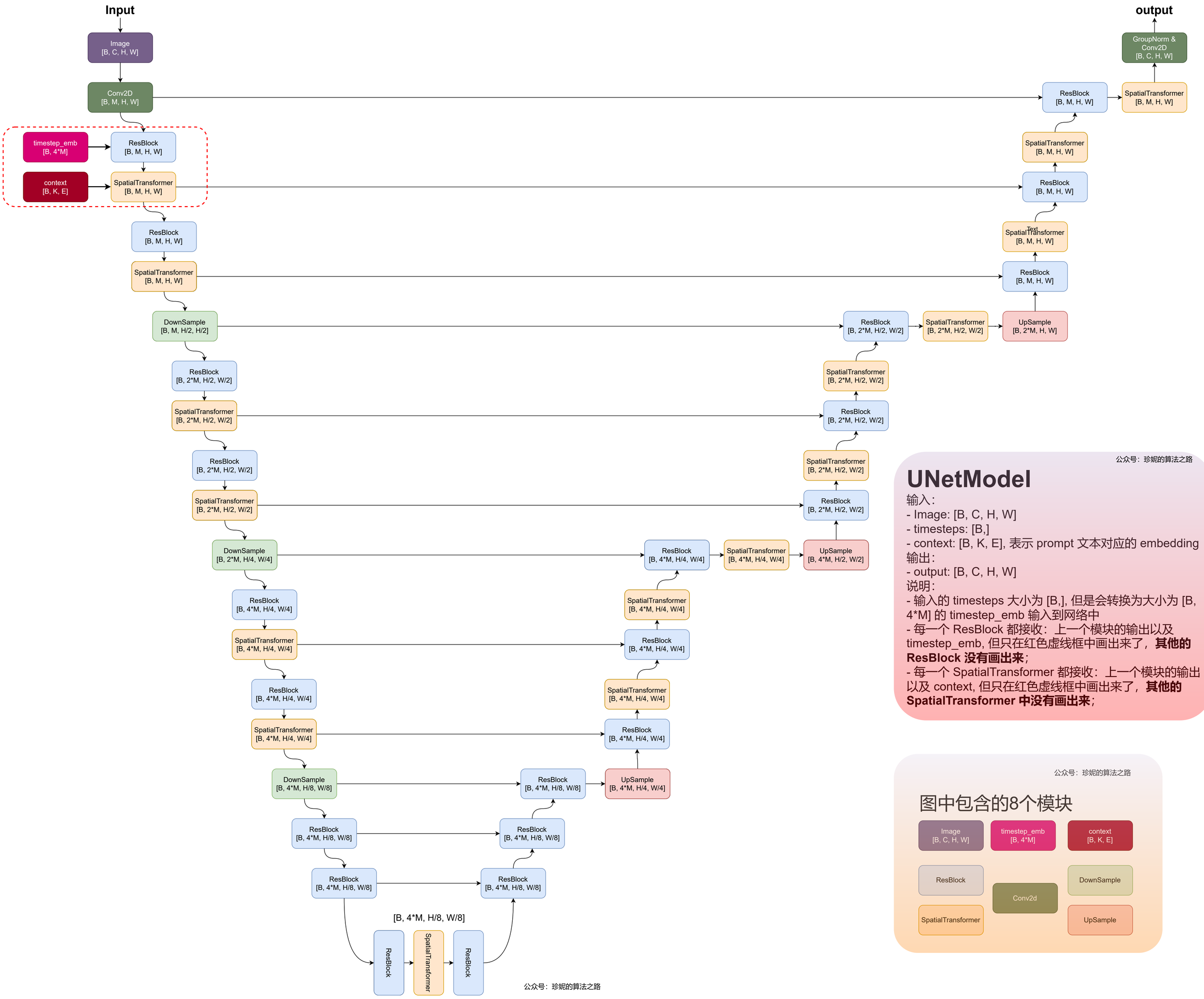

公众号：珍妮的算法之路

UNetModel

输入：
- Image: [B, C, H, W]
- timesteps: [B,]
- context: [B, K, E], 表示 prompt 文本对应的 embedding

输出：
- output: [B, C, H, W]

说明：
- 输入的 timesteps 大小为 [B,], 但是会转换为大小为 [B, 4*M] 的 timestep_emb 输入到网络中
- 每一个 ResBlock 都接收：上一个模块的输出以及 timestep_emb, 但只在红色虚线框中画出来了，**其他的 ResBlock 没有画出来**；
- 每一个 SpatialTransformer 都接收：上一个模块的输出以及 context, 但只在红色虚线框中画出来了，**其他的 SpatialTransformer 中没有画出来**；

公众号：珍妮的算法之路

图中包含的8个模块