对 Swin-Transformer 的补充总结

设计模块或者改进网络的本质就是人为地引入先验知识，即从人类如何理解这个任务或者图像的机制上，设计一个什么样的模块或者方法，让网络模拟人的这种机制去分析图像。

以 Swin 为例，原本将图像分为一个个 token 的办法是很呆板的，很可能刚好分在一个目标的中间，导致这个目标在不同的块上，于是我们认为这很影响模型的效果，然后 Swin 就设计了不同尺寸的窗口，避免了这样的问题。

又比如说，在目标检测中，有很大的目标也有很小的目标，当进行了五次降采样后，小目标很可能就不见了，大目标还在。早期的目标检测就是直接在最后一层预测所有的目标，然后发现这种方法对小目标不适用，为了解决这个问题，又衍生出在前面几层预测小目标，在后面几层预测大目标，但是这样又产生了新的问题，在前面几层预测小目标的时候，也有大目标物体的存在，但是却选择不预测，而在预测大目标的层上仍然有小目标，却只选择预测大目标，忽略小目标，所以就有人提出了 ASFF。

再比如说，图像增强的应用，我们选择使用神经网络的目的就是想让它可以灵活地应对各种情况，例如车头的识别，总不能训练的网络模型只能识别所有朝左的车头吧，所以设计了让它上下左右翻转的。

所以，设计模型首先最重要的就是分析数据，看看数据有什么特点，然后选择合适的网络，从得到的结果中分析网络存在哪些特点，例如分块比较呆板，例如目标在不同层的预测，这些导致了什么样的问题，这都是我们人类大脑对于识别或检测这个任务应该要关注哪些东西而给模型设计这样或者那样的模块或者方法。

相比于 CNNs 来说，Transformer 引入的归纳偏置更少，意味着所带来的约束就更少，相比于 CNN 的解空间就更大，更难收敛。同理，原本一个预测直线的任务只需要预测直线与标签之间的距离损失就可以，给它加上一个斜率损失就是给他增加了一个约束，让它可以更好地收敛，这个约束的本质就是人为引入的先验知识。