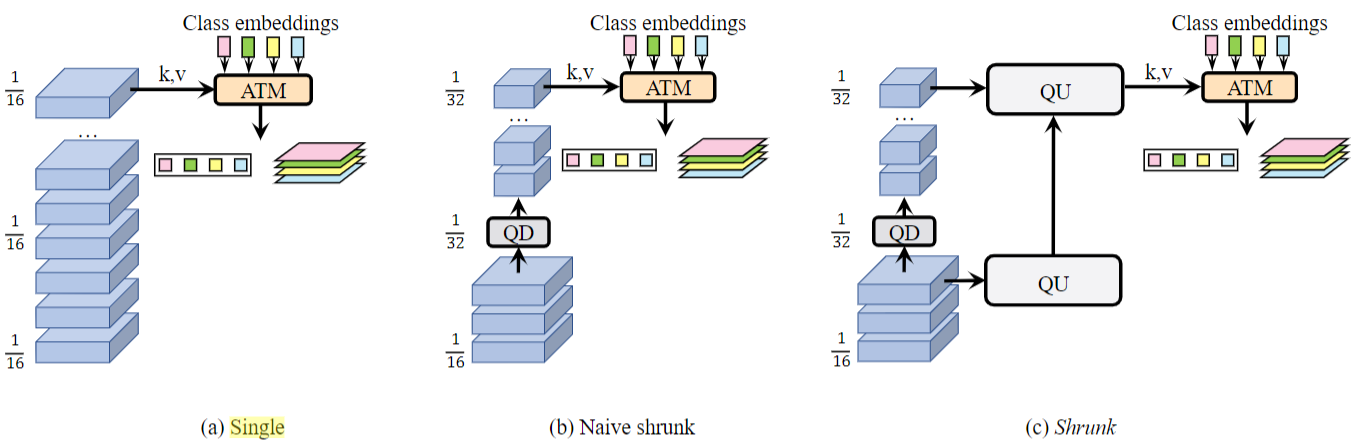


选择了vit的第8，16，24层的feature作为key和value，随机初始化的NXC的类嵌入为query，从高分辨率到低分辨率去attend query。选择的每一层vit 的feature会和query经过一个ATM模块，对应3层transformer decoder。其中一个ATM模块会输出一个mask，mask由ATM模块第三层transfromer decoder的cross attention图经过sigmoid得到，并且会和上面一个ATM模块输出的mask相加融合。因为使用 Softmax 函数会专门将注意力集中在相似度最大的token上，除了产生最大相似性的token之外，其他token也有意义。基于这种直觉，segvit设计了一个轻量级模块，可以更直接地生成语义预测。用最后一个ATM模块输出的query接FC后得到的class predictions（NXN），和最后一个ATM模块输出的mask predictions（NXHW）相乘得到预测的mask。



由于ViT 作为主干具有大的计算成本。论文提出了一种使用基于query的下采样 (QD) 和上采样 (QU) 的 Shrunk 结构。与上述原始模型不同的是，只选择了第8，24层的feature。

QD是会将高分辨率的feature下采样后的feature作为query。去attend 高分辨率的feature，从而得到被更新后的下采样后的feature。

相比a图使用vit作为backbone，计算量大。b图中，在主干的 1/3 深度（具有 24 层的主干的第 8 层）应用 QD一次，以将层的输出的分辨率从 1/16 下采样到 1/32，以降低整体计算成本。由于 QD 过程让信息丢失了，因此性能下降。c图对应的代码实现是，首先backbone上需要在第八层对该层输出feature上采样，将其输出作为query，1/16分辨率的feature作为key和value，得到被更新后的1/32分辨率的结果作为backbone该层应当输出的feature。再随机初始化1/16分辨率的query，去attend 高分辨率到低分率的backbone输出的feature。将被更新后的1/16的feature再作为key 和value，让随机初始化的NXC的token作为query去attend，也就是ATM结构，只不过shrunk版的只有一个ATM模块。

c图中，第一个 QU 层从主干的低层网络中获取分辨率为 1/16 的特征作为query，与来自主干最后一层的1/32分辨率的下采样特征进行交叉注意。该 QU 结构输出的形状分辨率为 1/16。

**结果：**

SegVit的在ade20k上的结果：

'†' means the models use the backbone weights pre-trained by AugReg [ 30 ]. '\*' represents the model is reproduced under the same settings as the official repo.

