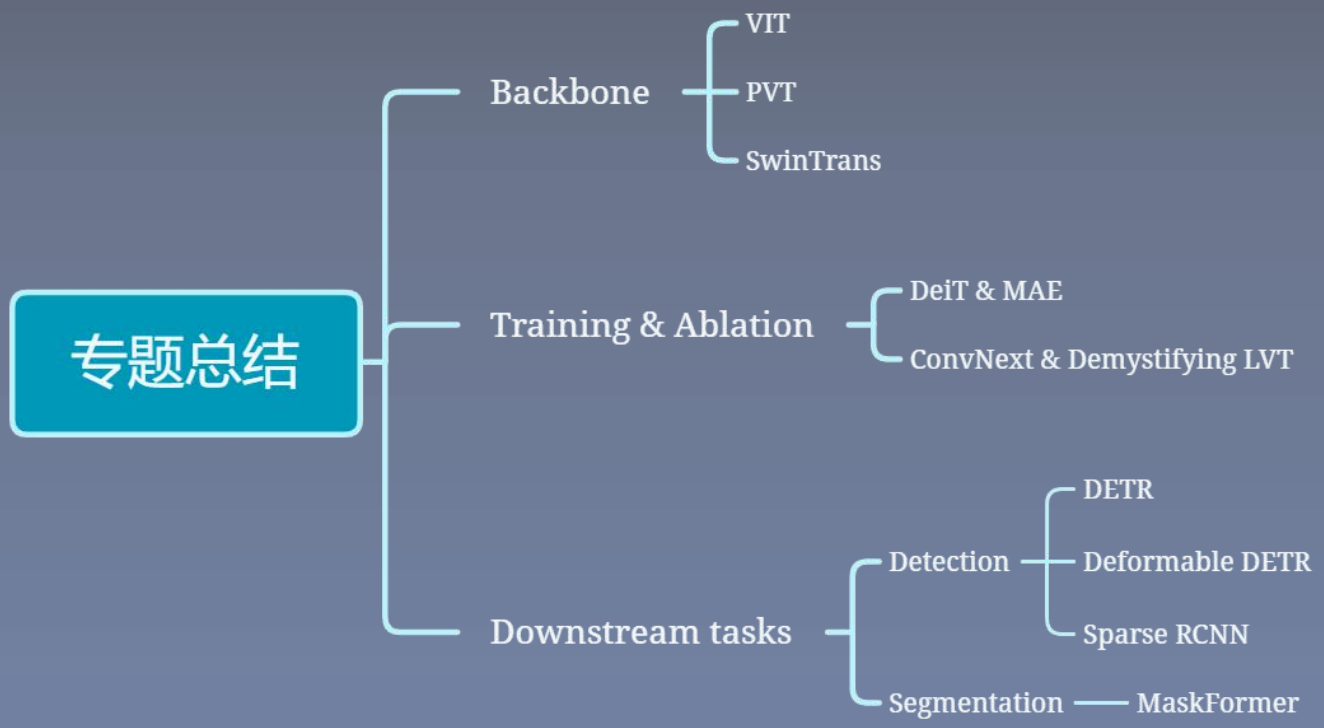
CV中的Transformer[9]-专题总结

导师：电子羊

专题总结

专题分为3个部分，9次课程，在这九次课程中我们精讲了下面的9~12篇论文



专题总结

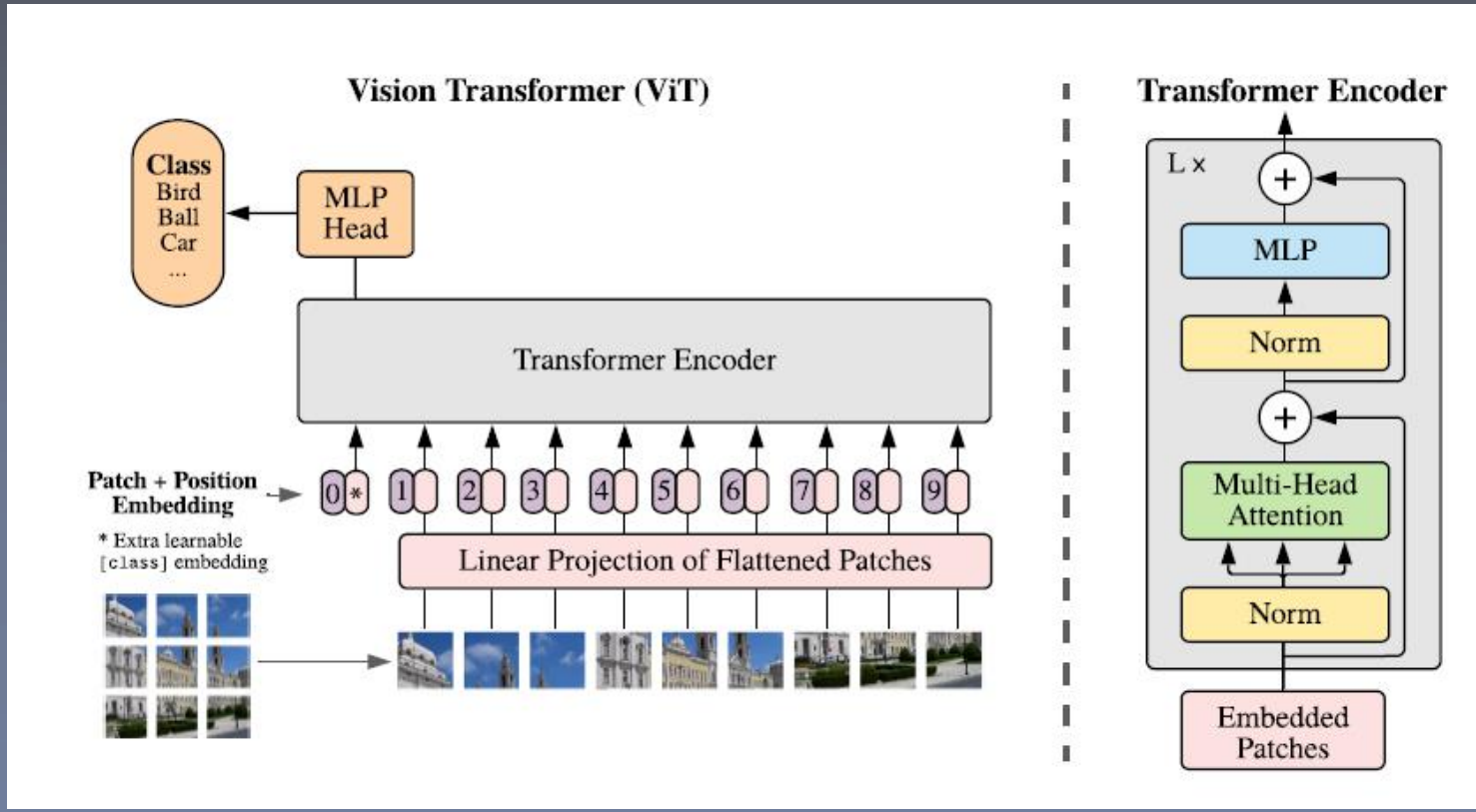
学了什么-Backbone部分

ViT

初见Transformer

论文课上详细的分析了Attention的原理以及它在计算上的优势

代码课上讲述了其代码实现



专题总结

学了什么-Backbone部分

PVT

下游任务可用的Transformer

主要修改了整体结构，从直筒型变为多阶段型

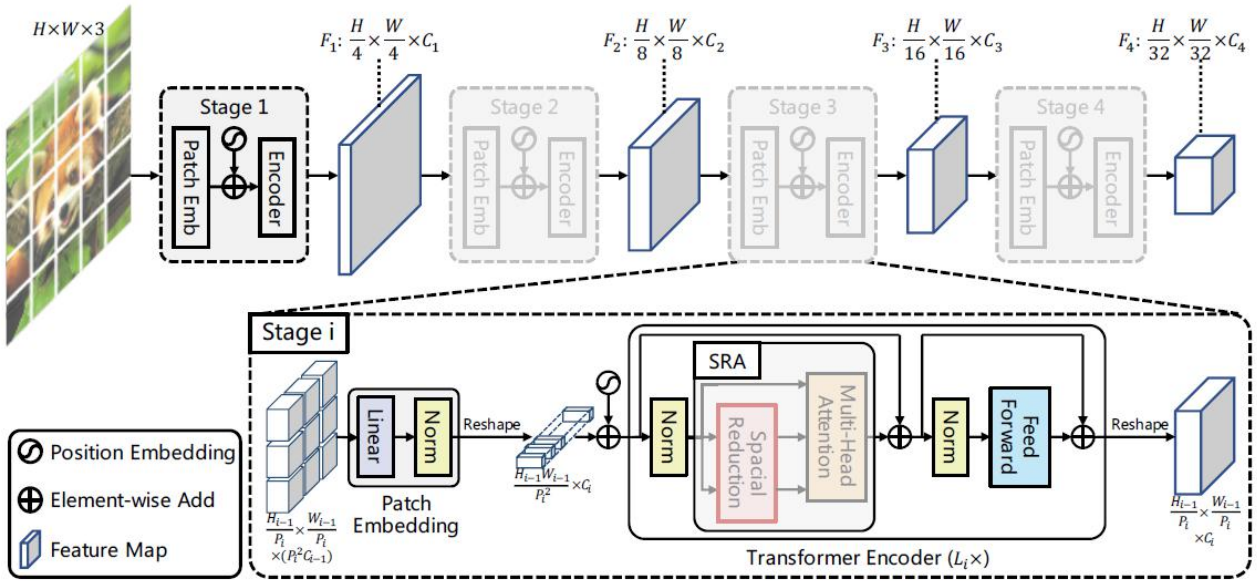


Figure 3: **Overall architecture of the proposed Pyramid Vision Transformer (PVT).** The entire model is divided into four stages, and each stage is comprised of a patch embedding layer, and a L_i-layer Transformer encoder. Following the pyramid structure, the output resolution of the four stages progressively shrinks from 4-stride to 32-stride.

专题总结

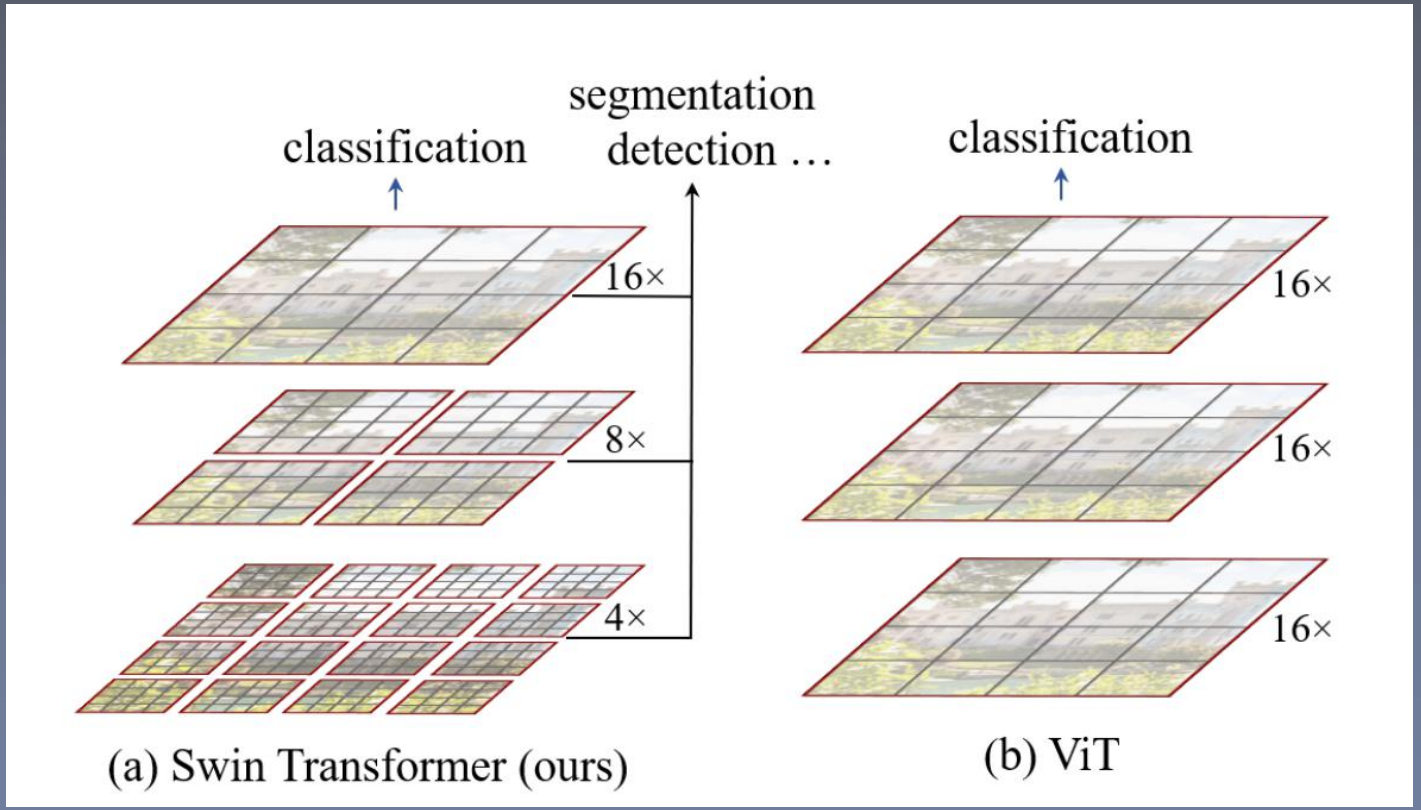
学了什么-Backbone部分

Swin Transformer

下游任务可用的Transformer

主要设定为MSA Local化

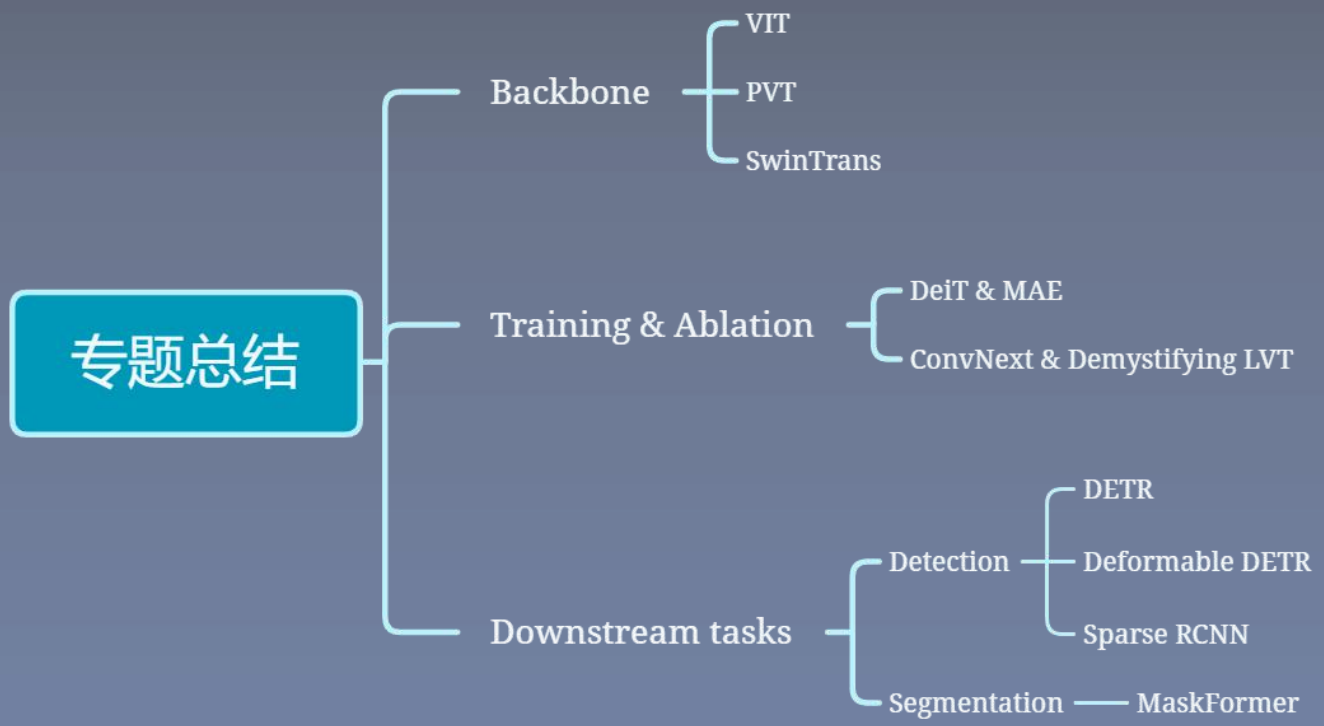
为解决MSA计算复杂度过高的问题
除了像PVT那样限制K的数量，还可以
直接把计算限制在一定的区域内。



专题总结

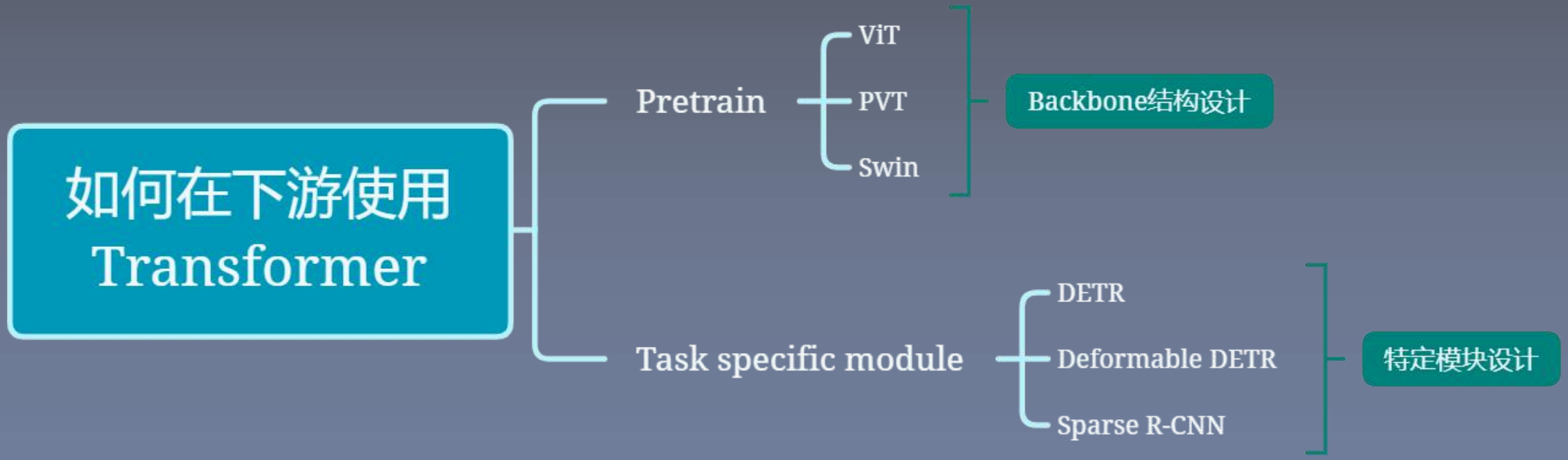
学了什么-Backbone部分

Backbone部分主要对应Transformer Encoder
优势为更强的特征提取能力，劣势为超大显存开销



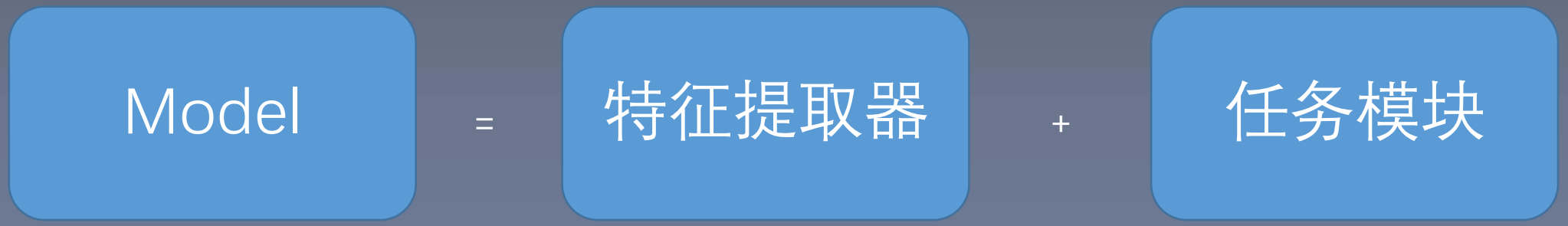
专题总结

学了什么-训练



专题总结

学了什么-训练



专题总结

学了什么-训练

DeiT

DeiT：一组训练VIT的新超参

Methods	ViT-B [15]	DeiT-B
Epochs	300	300
Batch size	4096	1024
Optimizer	AdamW	AdamW
learning rate	0.003	$0.0005 \times \frac{\text{batchsize}}{512}$
Learning rate decay	cosine	cosine
Weight decay	0.3	0.05
Warmup epochs	3.4	5
Label smoothing ε	✗	0.1
Dropout	0.1	✗
Stoch. Depth	✗	0.1
Repeated Aug	✗	✓
Gradient Clip.	✓	✗
Rand Augment	✗	9/0.5
Mixup prob.	✗	0.8
Cutmix prob.	✗	1.0
Erasing prob.	✗	0.25

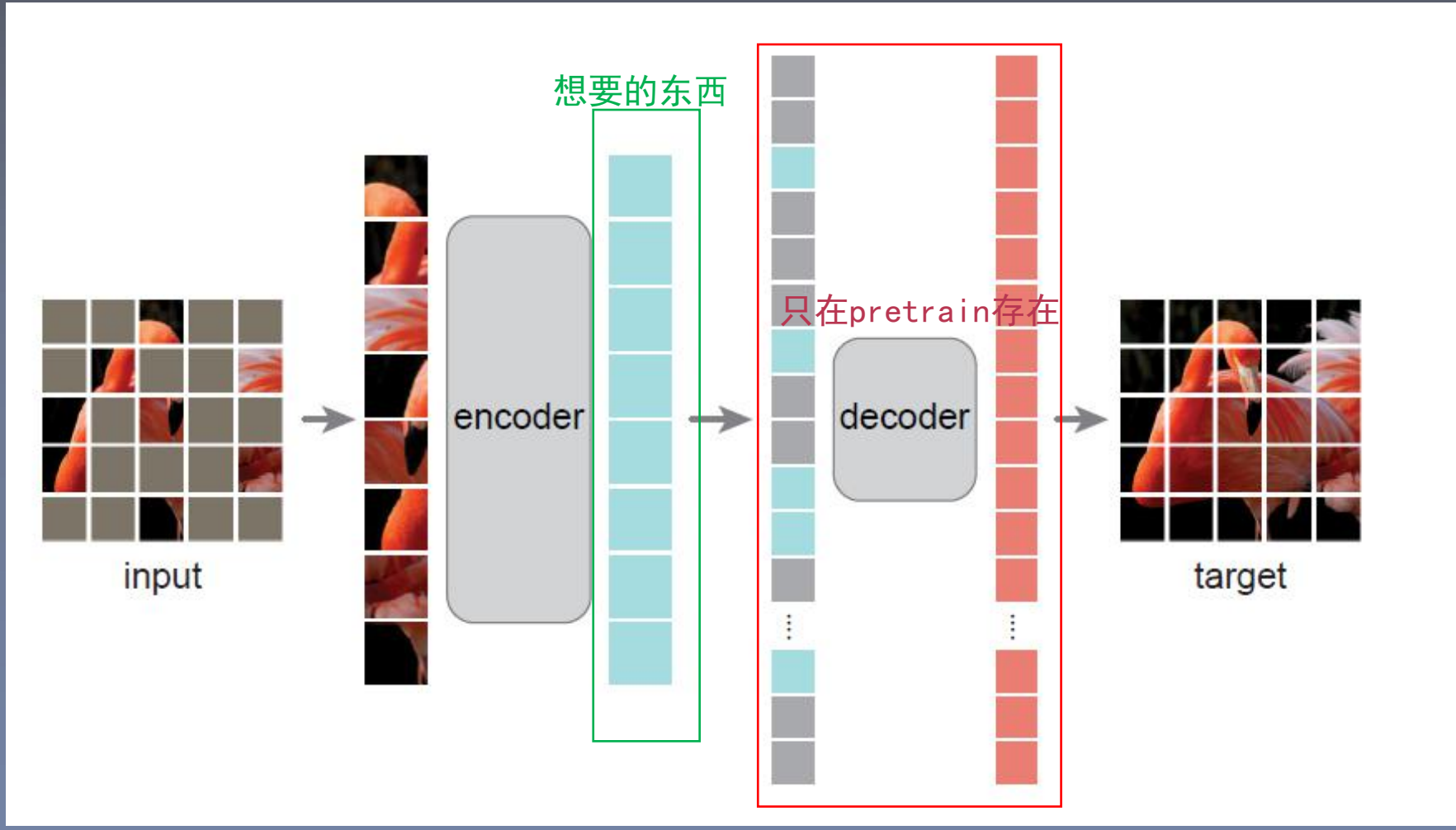
Table 9: Ingredients and hyper-parameters for our method and Vit-B.

专题总结

学了什么-训练

MAE

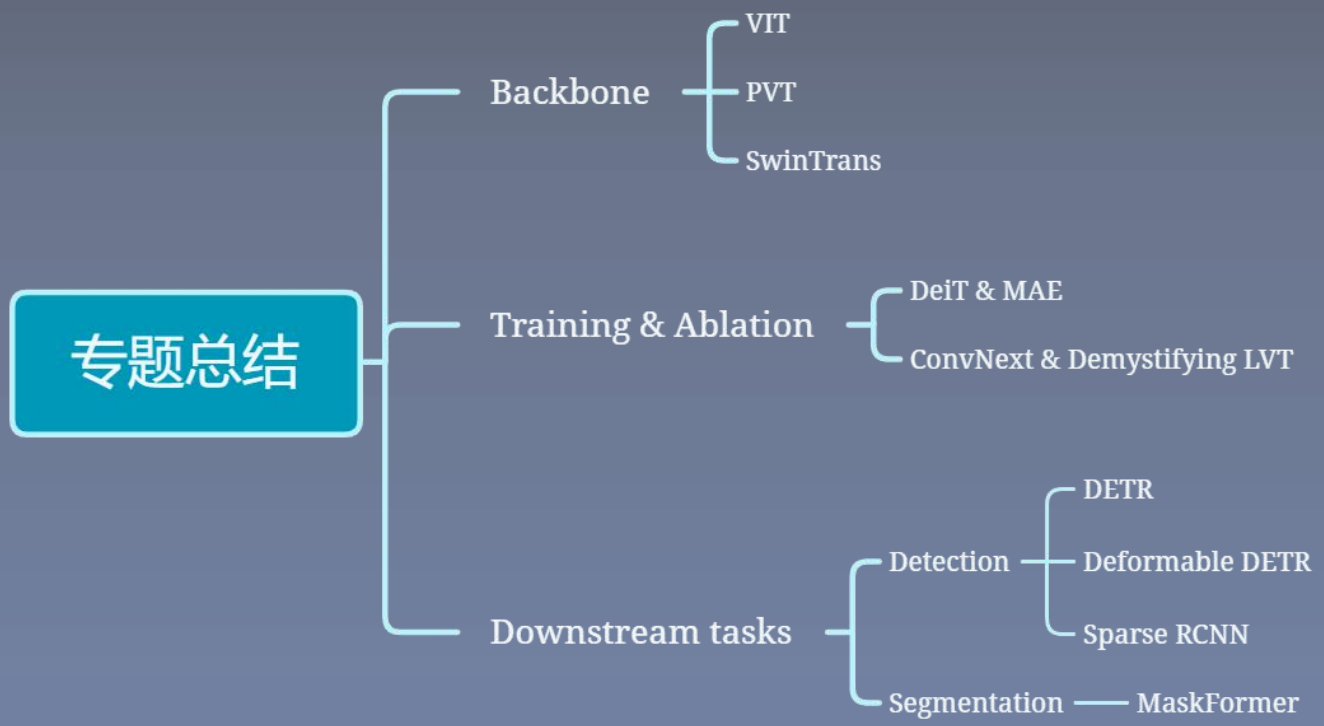
MAE：一种训练VIT的新任务



专题总结

学了什么-训练部分

训练部分主要讨论如何获得一组更好的参数
这部分主要对比了分类和重建两个任务



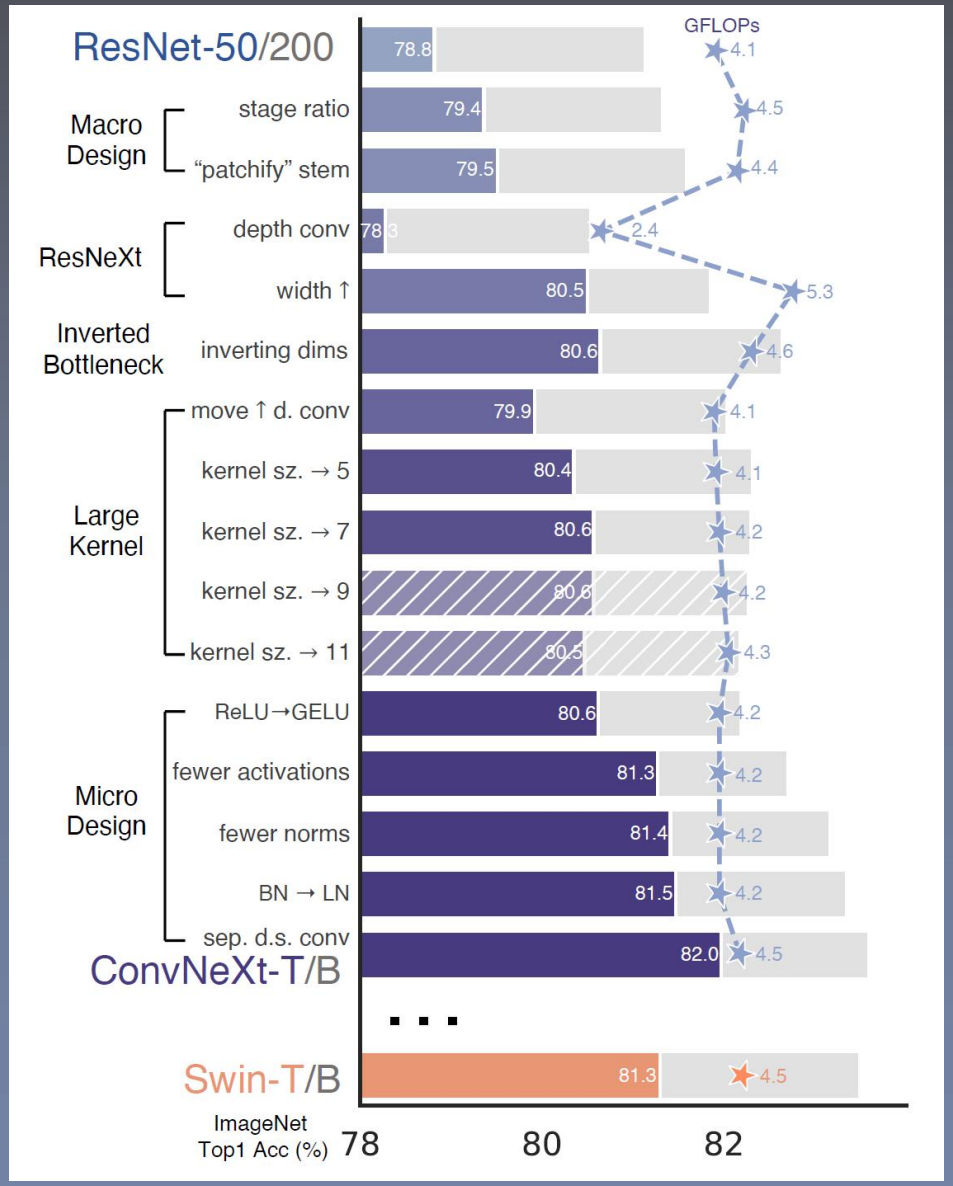
专题总结

学了什么-Ablation部分

ConvNext

20年代的CNNs

主要借鉴了Swin Transformer中的各种参数



专题总结

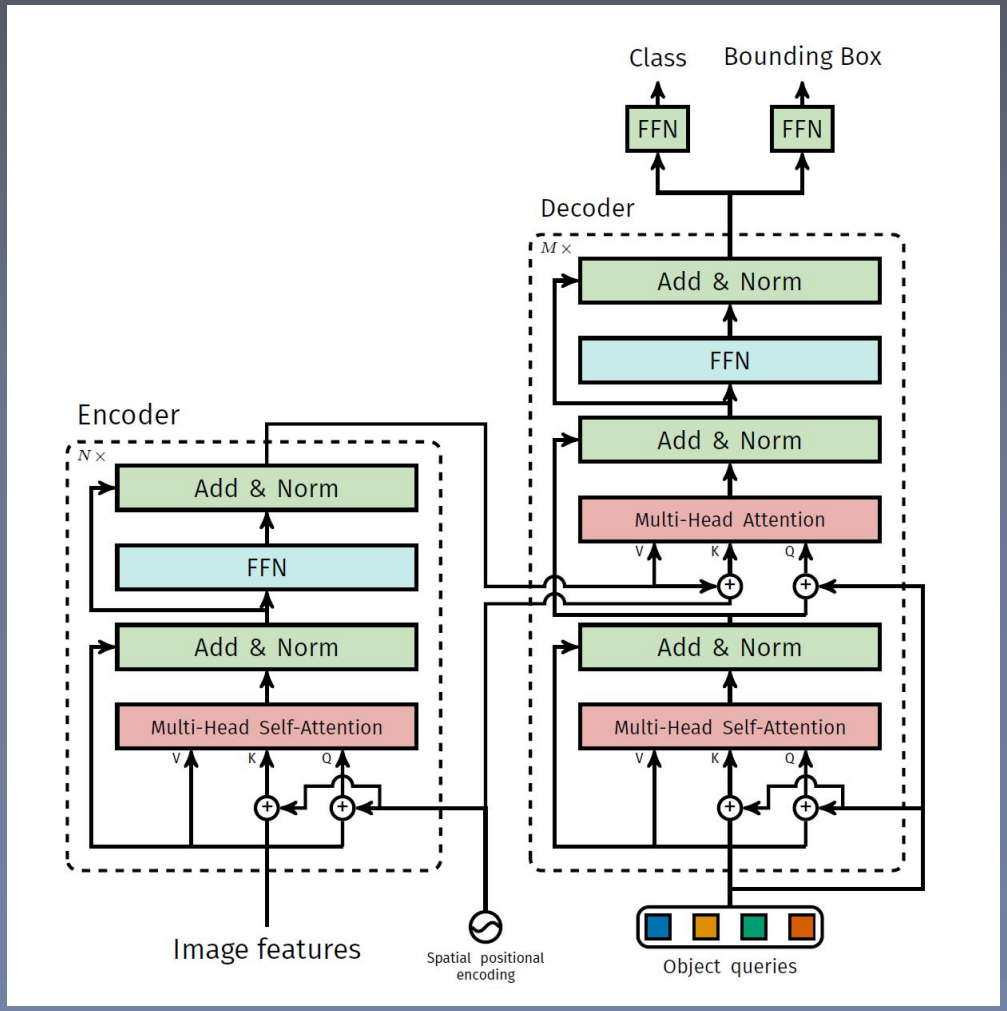
学了什么-下游任务

DETR

一对一的比对可以消除空间平滑->去NMS

讲解了一直以来目标检测的痛点->手工痕迹
Transformer Decoder初见

Object Query的作用？



专题总结

学了什么-下游任务

Deformable DETR

DCN与MSA的区别及联系

DETR收敛慢的原因

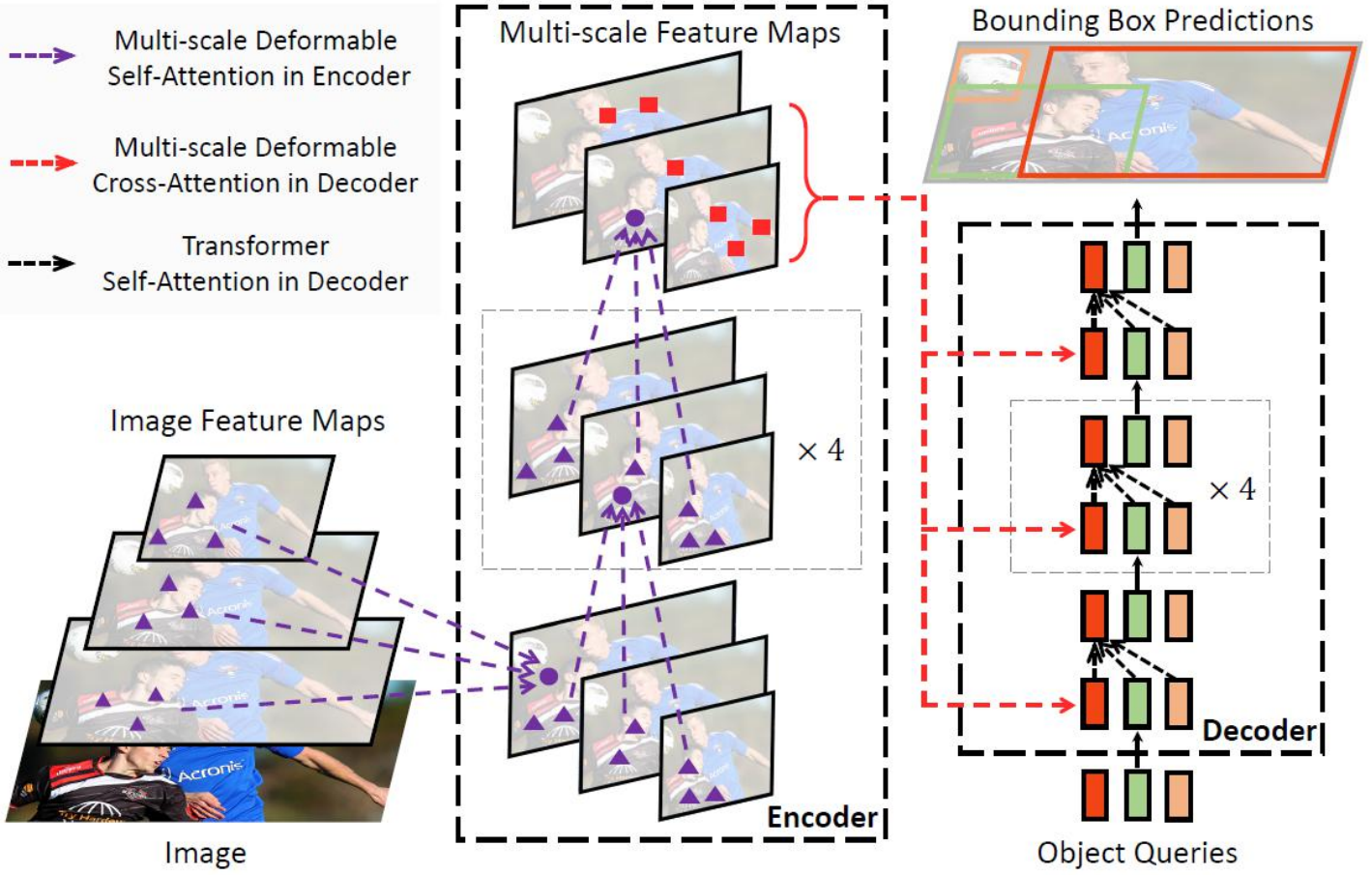


Figure 1: Illustration of the proposed Deformable DETR object detector.

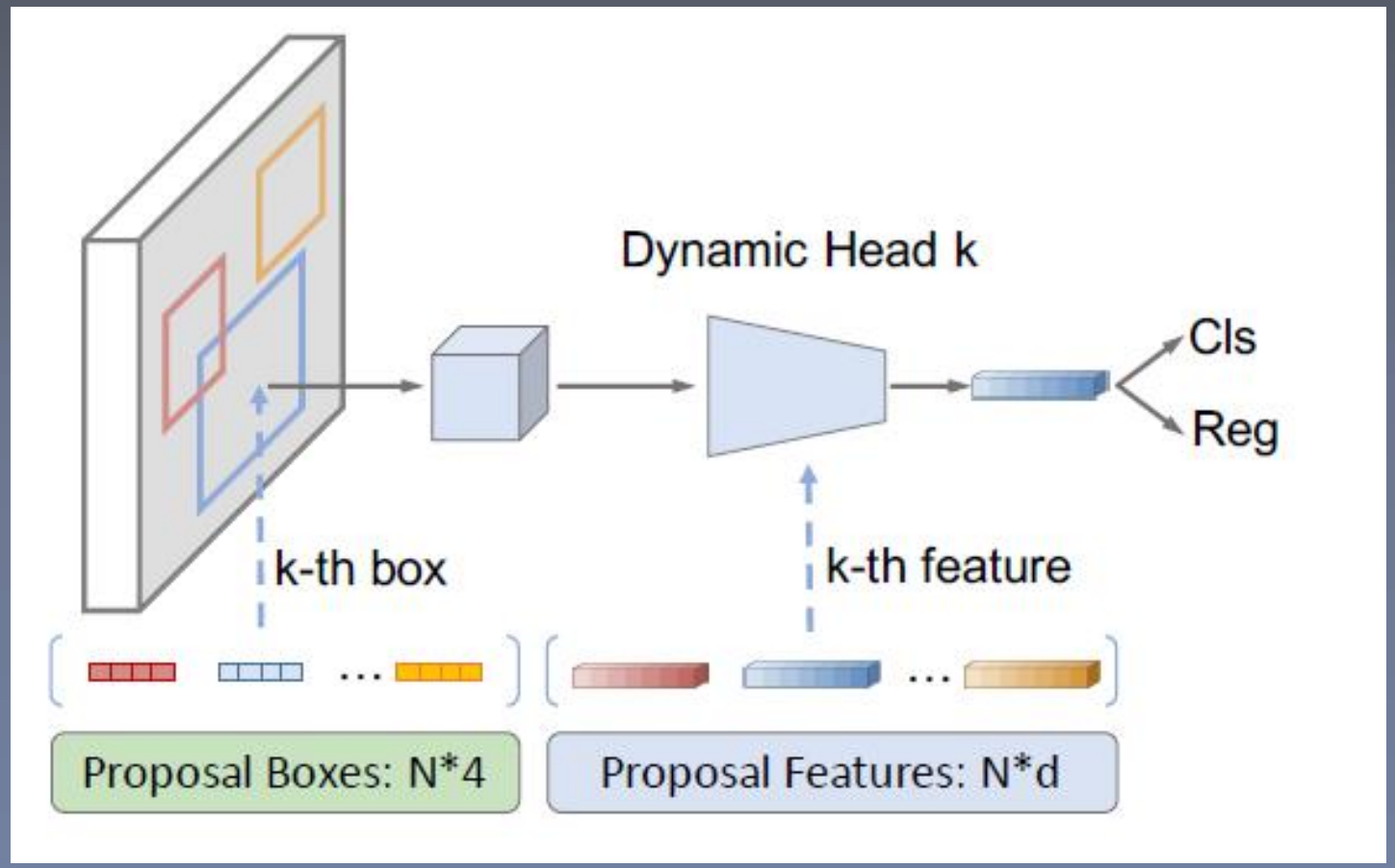
专题总结

学了什么-下游任务

Sparse RCNN

Dynamic Convs与MSA

纯Sparse结构



专题总结

学了什么-下游任务

MaskFormer

Mask Classification

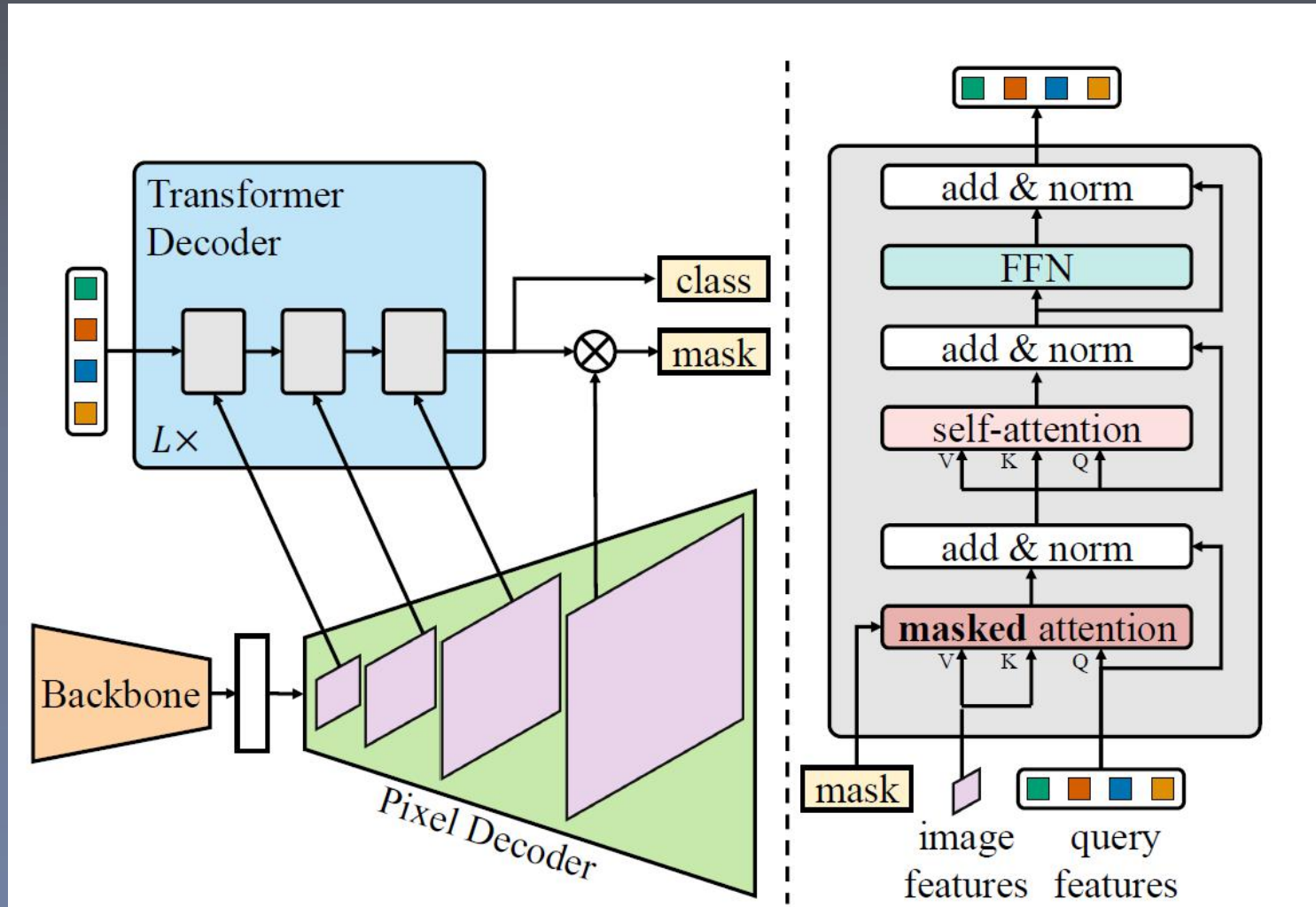


Figure 2. **Mask2Former overview.** Mask2Former adopts the

专题总结

学了什么

Transformer的优势

1. 有捕获Global信息的能力
2. 基于矩阵乘法，容易并行
3. 没有很强的Inductive Bias，拟合能力强
4. 点对点的Attention计算，没有空间平滑

Transformer的劣势

1. MSA的计算量过大
2. 显存开销高，无法用FPN，下游任务差
3. 没有Inductive Bias，需要更多的数据
4. Attention Map需要很长的iteration才能收敛

可能的改良

1. 在保留MSA的前提下，限制Q，K的数量(Swin, PVT都这么做的)
2. 用其他Attention的计算代替MSA，(Mixer, Deformable DETR)

专题总结

目标达成了吗？

专题学完获得什么？

- 深入理解Transformer的原理 √
- 实现过列出论文的代码 √
- 对CV中Transformer的优劣有自己的认识 √
- 能够根据理解做出自己的魔改 ？