Panoptic SegFormer

Zhiqi Li[1], Wenhai Wang[1], Enze Xie[2], Zhiding Yu[3],
Anima Anandkumar[3,4], Jose M. Alvarez[3], Tong Lu[1], Ping Luo[2]

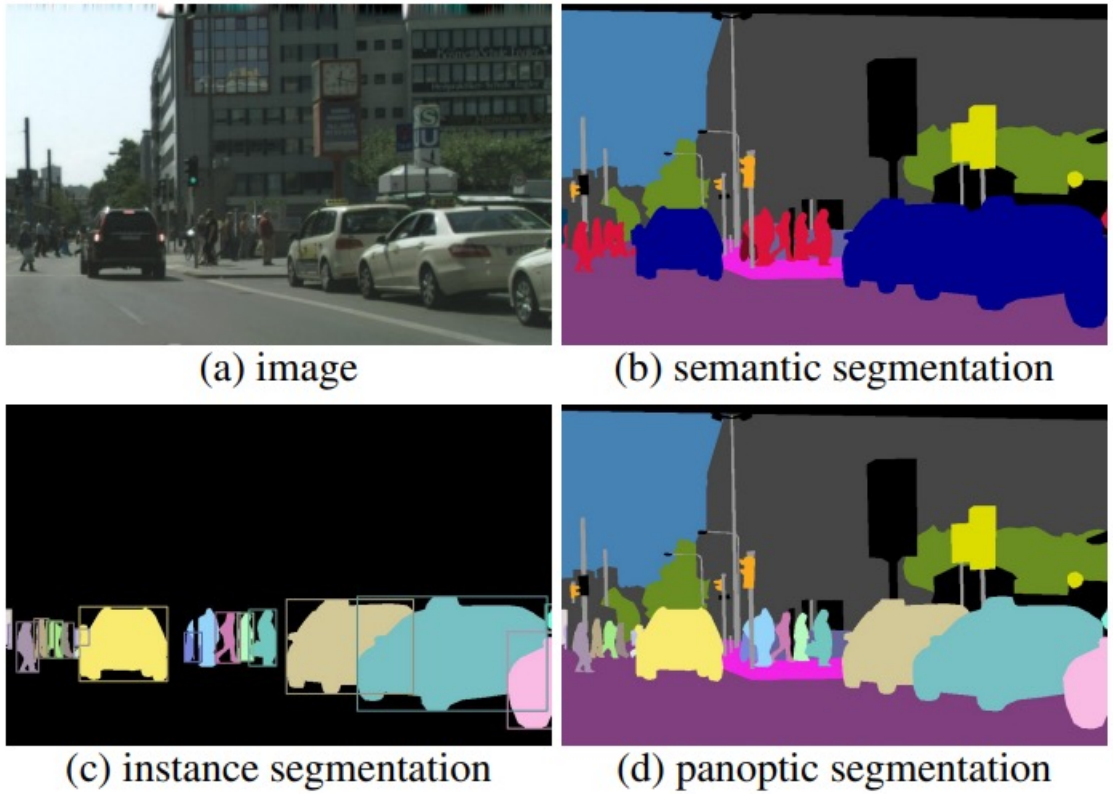
[1]Nanjing University [2]The University of Hong Kong [3]NVIDIA [4]Caltech

[2109.03814] Panoptic SegFormer (arxiv.org)

Panoptic Segmentation

Things: person, car, and bicycle

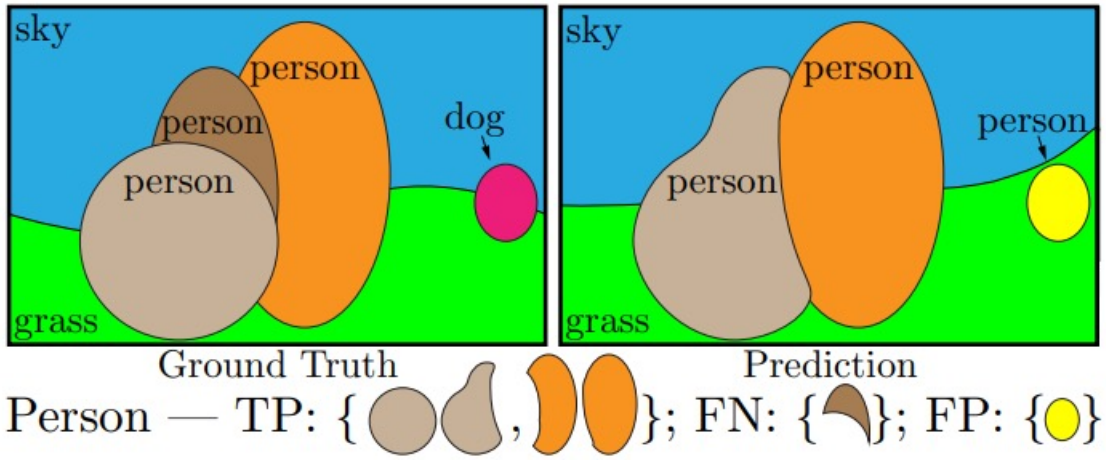
Stuff: sky, grassland, and snow



(a) image (b) semantic segmentation

(c) instance segmentation (d) panoptic segmentation

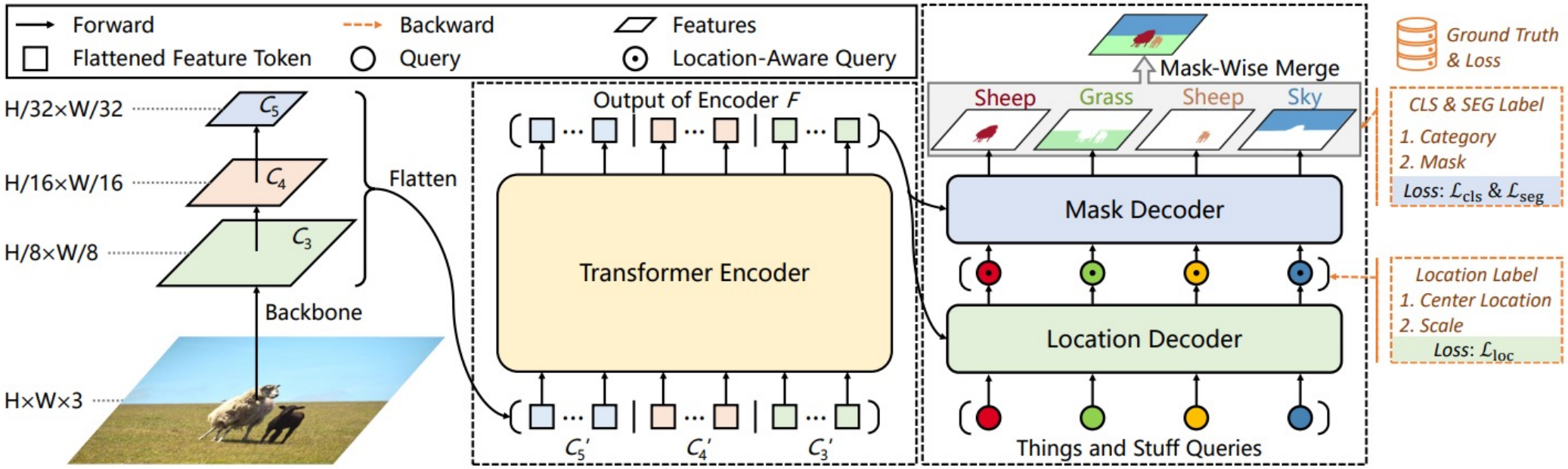
Panoptic Quality



$$\mathrm{PQ} = \frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}.$$

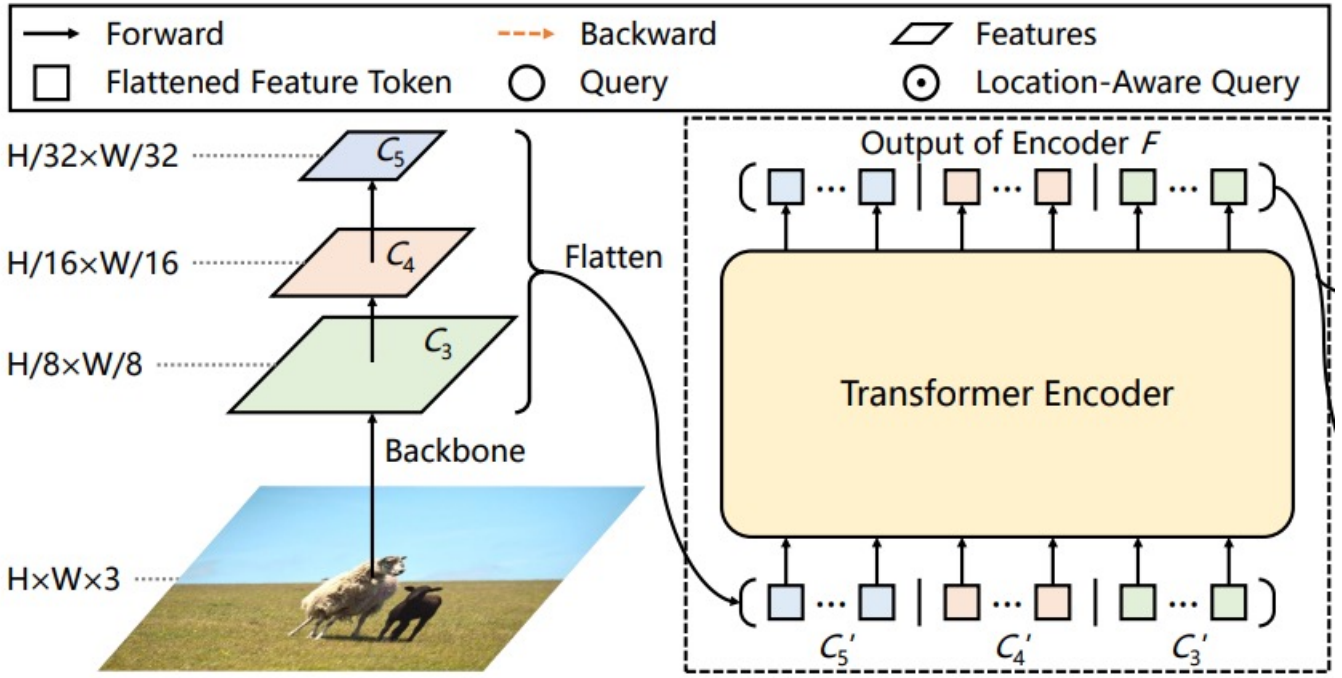
$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}}.$$

Overall architecture



Transformer Encoder

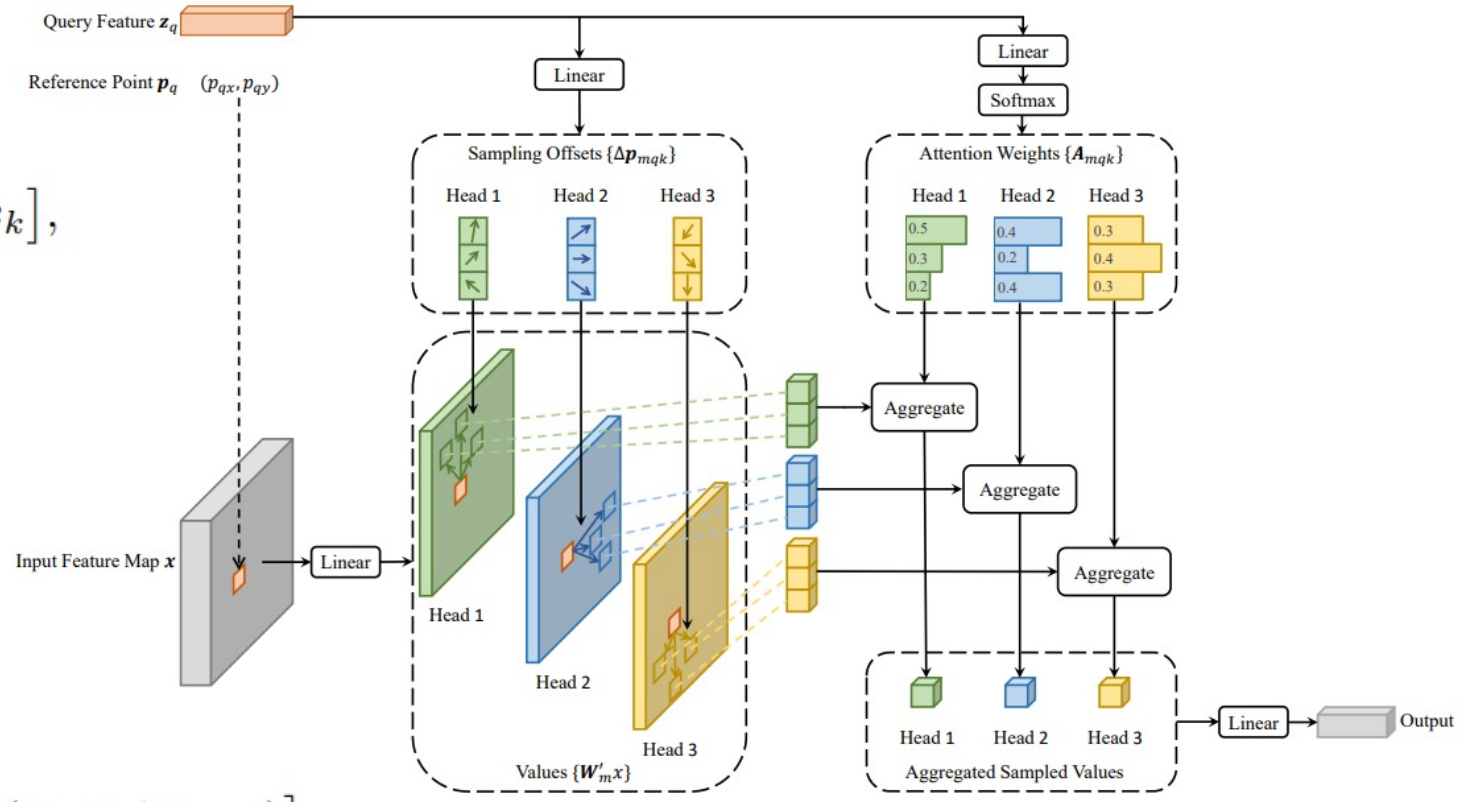
$C_3', C_4', C_5' \in R^{L_1 \times 256}, R^{L_2 \times 256}, R^{L_3 \times 256}$



Deformable Attention

https://arxiv.org/abs/2010.04159

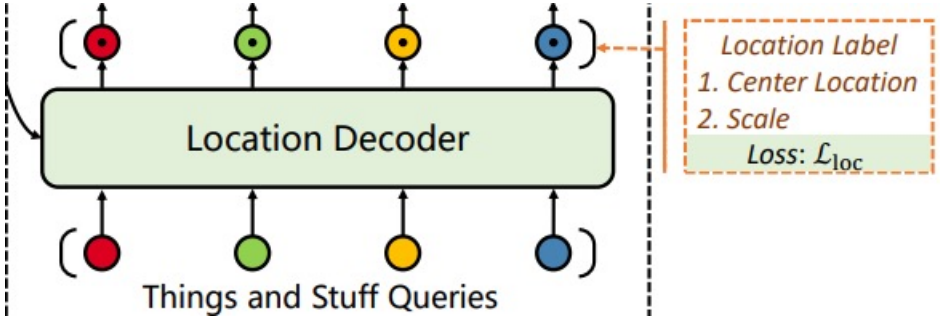
$$\text{MultiHeadAttn}(\boldsymbol{z}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k \in \Omega_k} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}_k \big],$$



$$\text{DeformAttn}(\boldsymbol{z}_q, \boldsymbol{p}_q, \boldsymbol{x}) = \sum_{m=1}^{M} \boldsymbol{W}_m \big[\sum_{k=1}^{K} A_{mqk} \cdot \boldsymbol{W}'_m \boldsymbol{x}(\boldsymbol{p}_q + \Delta \boldsymbol{p}_{mqk}) \big],$$

Location Decoder

- Introduce the location information (i.e., center location and scale) of things and stuff into the learnable queries.
- Given N randomly initialized queries
- Apply an auxiliary MLP head on top of location-aware queries to predict the center locations and scales of the target object



$$\mathcal{L}_{\text{loc}} = \sum_{i}^{N} \mathbb{1}_{\{y_i \neq \varnothing\}} \left(\mathcal{L}_1(f_c(m_i), \hat{u}_{\sigma(i)}) + \mathcal{L}_1(f_s(m_i), \hat{v}_{\sigma(i)}) \right),$$

Mask Decoder

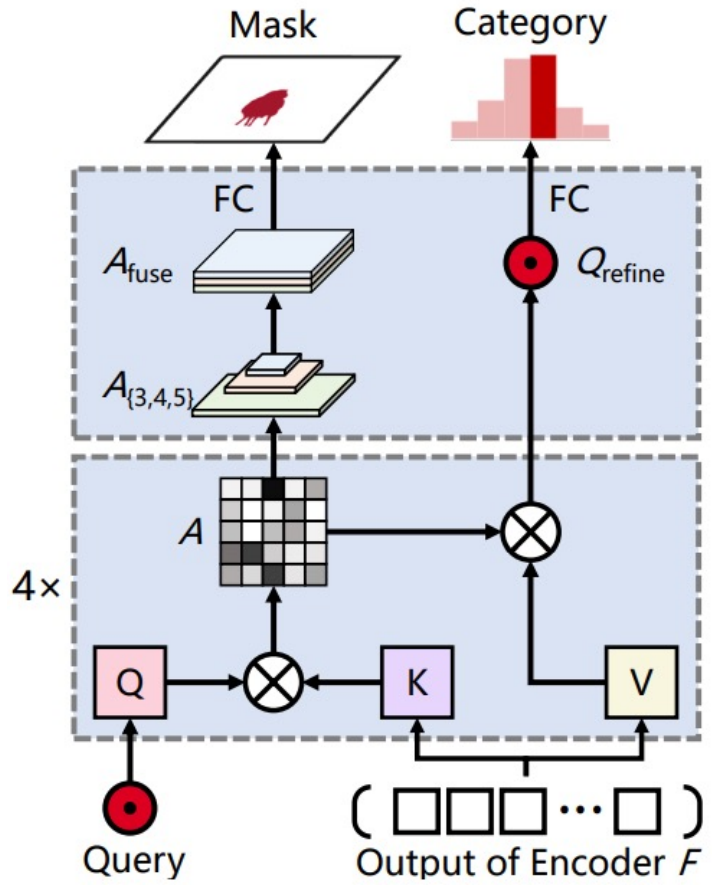
$$A \in \mathbb{R}^{N \times h \times (L_1+L_2+L_3)} \qquad Q_{\text{refine}} \in \mathbb{R}^{N \times 256}$$

$$(A_3, A_4, A_5) = \text{Split}(A), \quad A_i \in \mathbb{R}^{\frac{H}{2^{i+2}} \times \frac{W}{2^{i+2}} \times h},$$

Upsample these attention maps to the resolution of H/8×W/8 and concatenate them along the channel dimension

$$A_{\text{fuse}} = \text{Concat}(A_1, \text{Up}_{\times 2}(A_2), \text{Up}_{\times 4}(A_3)).$$

the mask decoder by the **common multi-head attention**



Loss Function

- the prediction set $\hat{Y} = \{\hat{y}_i\}_{i=1}^{N}$
- the ground truth set Y $= \{y_i\}_{i=1}^{M}$
- $N \geq M$

$$\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{loc}}\mathcal{L}_{\text{loc}}$$

$$\mathcal{L}_{\text{loc}} = \sum_{i}^{N} \mathbb{1}_{\{y_i \neq \varnothing\}} \left(\mathcal{L}_1(f_c(m_i), \hat{u}_{\sigma(i)}) + \mathcal{L}_1(f_s(m_i), \hat{v}_{\sigma(i)})\right),$$

Mask-Wise Inference

- *SemMsk* and *IdMsk* are first initialized by zeros
- Sorted prediction results in descending order of confidence score
- The results with confidence scores below thr_{cls} will be discarded
- The overlaps with lower confidence score will be removed

Algorithm 1: Mask-Wise Merging

```
def MaskWiseMergeing(c,s,m):
    # category c ∈ R^N
    # confidence score s ∈ R^N
    # mask m ∈ R^{N×H×W}
    SemMsk = np.zeros(H,W)
    IdMsk = np.zeros(H,W)
    order = np.argsort(-s)
    id = 0
    for i in order:
        # drop low quality results
        if s[i]< thr_cls:
            continue
        # drop overlaps
        m_i = m[i] & (SemMsk>0)
        SemMsk[m_i] = c[i]
        if isThing(c[i]):
            IdMsk[m_i] = id
            id += 1
    return SemMsk,IdMsk
```

Experiment

Method	Backbone	Epochs	PQ	PQ^{th}	PQ^{st}	#Param	FLOPs
Panoptic FPN [2]	R50-FPN [24, 39]	36	41.5	48.5	31.1	-	-
SOLOv2 [12]	R50-FPN	36	42.1	49.6	30.7	-	-
DETR [15]	R50	$\sim 150 + 25$	43.4	48.2	36.3	42.8M	137G
Panoptic FCN [13]	R50-FPN	36	43.6	49.3	35.0	37.0M	244G
K-Net [14]	R50-FPN	36	45.1	50.3	37.3	-	-
MaskFormer [17]	R50	300	46.5	51.0	39.8	45.0M	181G
DETR [15]	R101	$\sim 150 + 25$	45.1	50.5	37.0	61.8M	157G
Max-Deeplab-S [16]	Max-S	54	48.4	53.0	41.5	61.9M	162G
MaskFormer [17]	R101	300	47.6	52.5	40.3	64.0M	248G
Max-Deeplab-L [16]	Max-L	54	51.1	57.0	42.2	451.0M	1846G
MaskFormer [17]	Swin-L† [20]	300	52.7	58.5	44.0	212.0M	792G
Panoptic SegFormer	R50	12	46.4	52.6	37.0	47.0M	246G
Panoptic SegFormer	R50	50	50.0	56.1	40.8	47.0M	246G
Panoptic SegFormer	R101	50	50.4	56.3	41.6	65.9M	322G
Panoptic SegFormer	PVTv2-B0 [40]	50	49.6	55.5	40.6	22.2M	156G
Panoptic SegFormer	PVTv2-B2 [40]	50	52.6	58.7	43.3	41.6M	219G
Panoptic SegFormer	PVTv2-B5 [40]	50	54.1	60.4	44.6	100.9M	391G

Table 1: **Experiments on COCO val set.** Panotic SegFormer achieves 50.0% PQ on COCO `val` with ResNet-50 as backbone, surpasses previous methods such as DETR [15] and Panoptic FCN [17] over 6.6% PQ and 6.4% PQ respectively. Under training for 12 epochs, Panoptic SegFormer can achieve 46.4% PQ, which is comparable with 46.5% PQ of MaskFormer [17] that training for 300 epochs. † notes that backbones are pre-trained on ImageNet-22K.

Experiment

Method	Backbone	Epochs	PQ	PQ^{th}	PQ^{st}	#Param	FLOPs
Panoptic FPN [2]	R101-FPN	36	43.5	50.8	32.5	-	-
DETR [15]	R101	$\sim 150 + 25$	46.0	-	-	61.8M	157G
Panoptic FCN [13]	R101-FPN	36	45.5	51.4	36.4	56.0M	310G
K-Net [14]	R101-FPN	36	47.0	52.8	38.2	-	-
Max-Deeplab-S [16]	Max-S [16]	54	49.0	54.0	41.6	61.9M	162G
K-net [14]	Swin-L†	36	52.1	58.2	42.8	-	-
Max-Deeplab-L [16]	Max-L [16]	54	51.3	57.2	42.4	451.0M	1846G
Innovation [22]	ensemble	-	53.5	61.8	41.1	-	-
Panoptic SegFormer	R50	50	50.0	56.2	40.8	47.0M	246G
Panoptic SegFormer	R101	50	50.9	57.1	41.4	65.9M	322G
Panoptic SegFormer	PVTv2-B5 [40]	50	54.4	61.1	44.3	100.9M	391G

Table 2: **Experiments on COCO test-dev set.** With PVTv2-B5 [40] as backbone, Panoptic SegFormer achieves 54.4% PQ on COCO `test-dev`, surpassed previous SOTA methods Max-Deeplabe-L [16] and competition-level methods Innovation [22] over 3.1% PQ and 0.9% PQ respectively with fewer parameters and computation cost.

Experiment

Method	Backbone	Epochs	AP^{seg}	AP^{seg}_S	AP^{seg}_M	AP^{seg}_L
Mask R-CNN [29]	R50-FPN	36	37.5	21.1	39.6	48.3
SOLOv2 [12]	R50-FPN	36	38.8	16.5	41.7	56.2
SOLQ (300 queries) [31]	R50	50	39.7	21.5	42.5	53.1
HTC [41]	R50-FPN	36	40.1	23.3	42.1	52.0
QueryInst(300 queries) [32]	R50-FPN	36	40.6	**23.4**	42.5	52.8
Panoptic SegFormer (300 queries)	R50	50	**41.7**	21.9	**45.3**	**56.3**

Table 3: **Instance segmentation experiments on COCO `test-dev` set.** When training with things only, Panoptic SegFormer can perform instance segmentation. With ResNet-50 as backbone, Panoptic SegFormer achieves 41.7 mask AP on COCO `test-dev`, which is 1.6 AP higher than HTC [41].