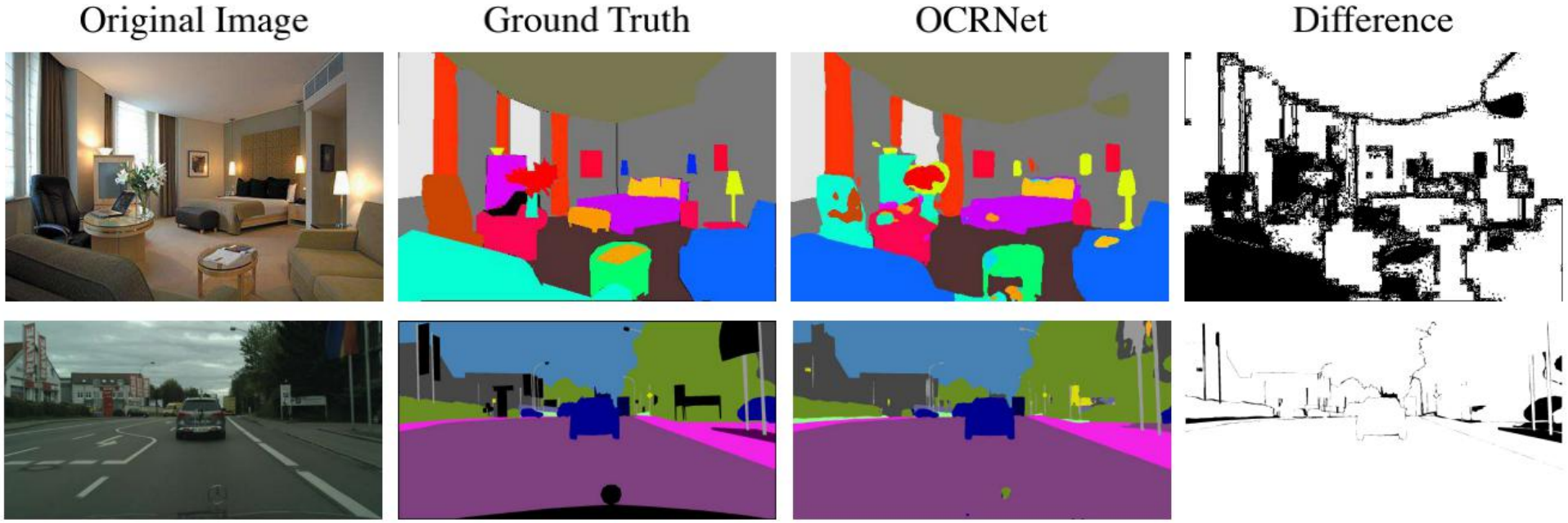
A Unified Efficient Pyramid Transformer for Semantic Segmentation

Fangrui Zhu*[1], Yi Zhu[2], Li Zhang[1], Chongruo Wu[3], Yanwei Fu[1], Mu Li[2]
[1] School of Data Science, Fudan University
[2] Amazon Web Services
[3] University of California, Davis

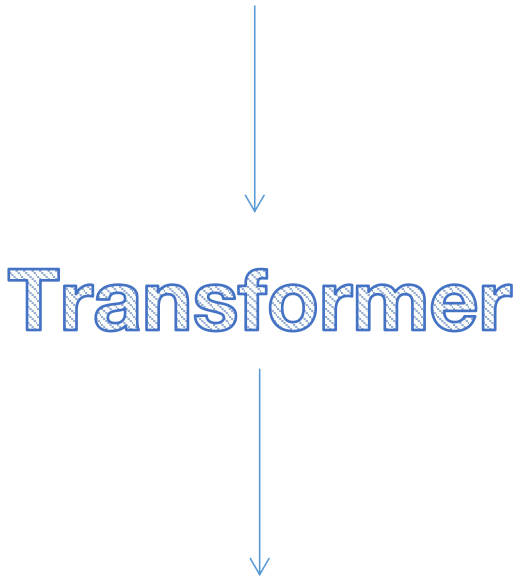
语义分割存在的广泛问题

- ADE20K 类别比较多的情况-> 类别混淆
- Cityscapes 类别不多 -> 边界混淆



解决策略

- 全局语义信息
- 局部细节



大图像带来的爆显存问题

Contributions

- 提出了新的attention模块--efficient pyramid transformer, 充分学习图像的上下文信息
- 引入空间分支，提供自适应图像特征信息以及边界信息
- 实验效果，在ADE20K，Cityscapes, Pascal-Context 表现卓越

Related work （context modeling && boundary handling）

- FCN
- pyramid pooling
- Global pooling
- attention mechanism(last convolutional features)

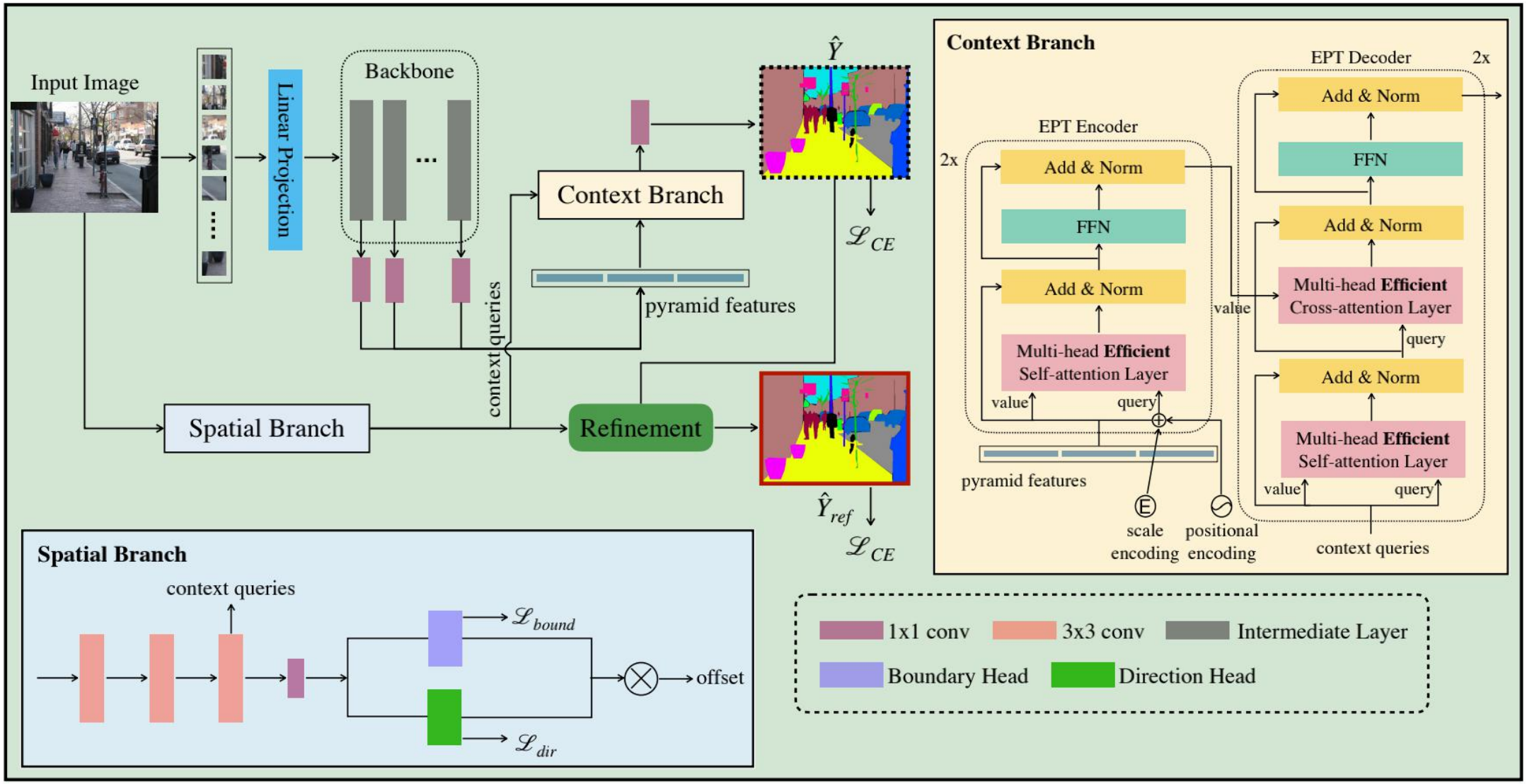
Transformer(sparse sampling strategy)

Related work（context modeling && boundary handling）

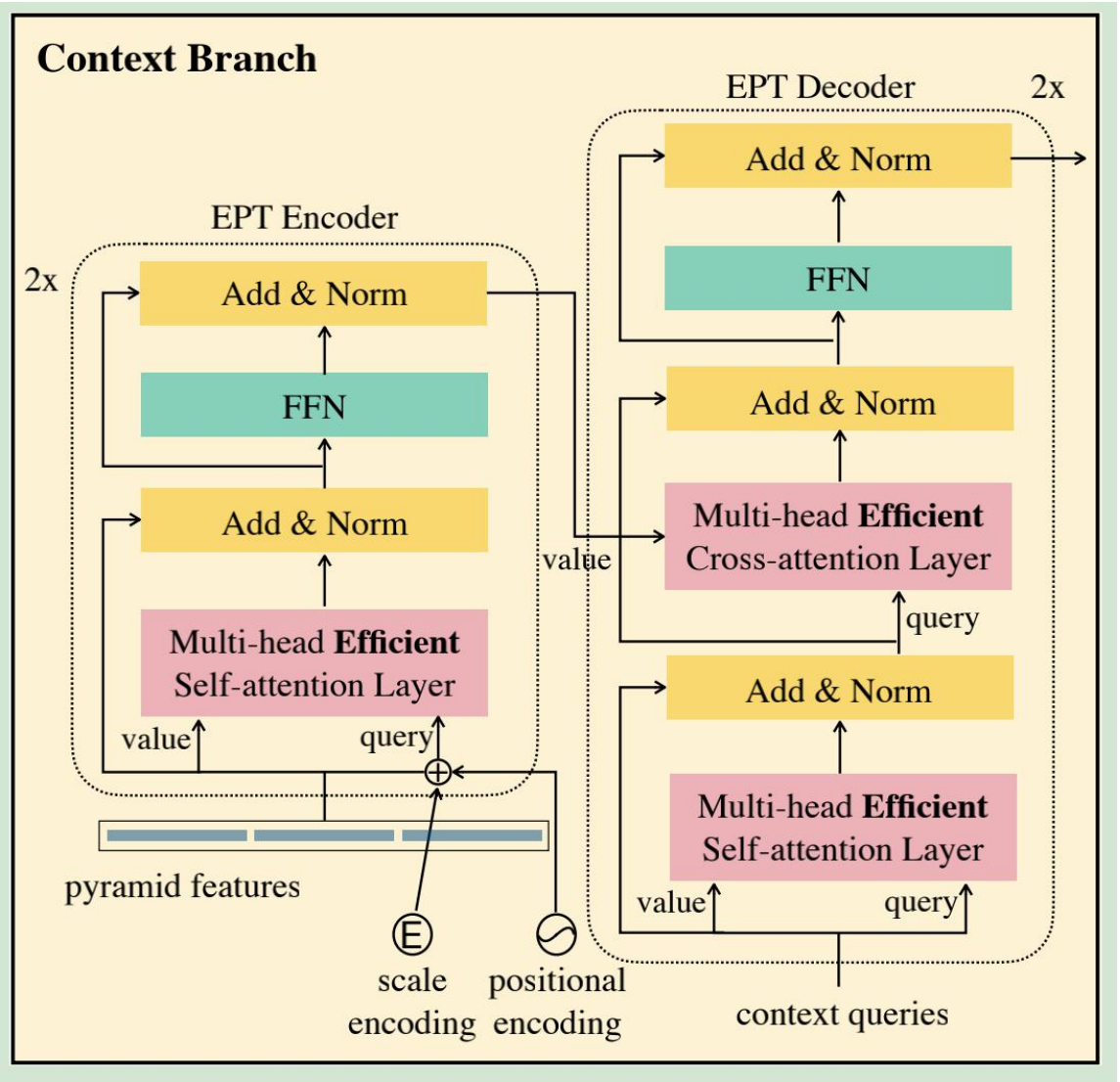
- localizing semantic boundaries
- refining boundary segmentation results

1. 没有考虑全局语义信息
2. 大部分工作是two-stage的

Method


Input Image
Linear Projection
Backbone
Context Branch
pyramid features
context queries
Spatial Branch
Refinement
$\hat{Y}$
$\mathcal{L}_{CE}$
$\hat{Y}_{ref}$
$\mathcal{L}_{CE}$
Context Branch
EPT Encoder
2x
Add & Norm
FFN
Add & Norm
Multi-head Efficient Self-attention Layer
value
query
pyramid features
scale encoding
positional encoding
EPT Decoder
2x
Add & Norm
FFN
Add & Norm
Multi-head Efficient Cross-attention Layer
query
Add & Norm
Multi-head Efficient Self-attention Layer
value
query
context queries
Spatial Branch
context queries
$\mathcal{L}_{bound}$
$\mathcal{L}_{dir}$
offset
1x1 conv
3x3 conv
Intermediate Layer
Boundary Head
Direction Head

Efficient Transformer for modeling contexts



1. 输入像素序列过长会爆显存

2. 对于任务本身而言也不需要输入全部像素

General MSA in transformer

$$K \in \mathbb{R}^{n \times d_m} \qquad Q \in \mathbb{R}^{n \times d_m} \qquad V \in \mathbb{R}^{n \times d_m}$$

$$A_m = \text{softmax}\left(\frac{QW_m^Q \left(KW_m^K\right)^{\mathrm{T}}}{\sqrt{d_{model}}}\right) \qquad W_m^Q, W_m^K \in \mathbb{R}^{d_m \times d_k} \qquad A_m \in \mathbb{R}^{n \times n}$$

$$\text{Attn}_m = A_m V W_m^V \qquad W_m^V \in \mathbb{R}^{d_m \times d_v} \qquad \text{Attn}_m \in \mathbb{R}^{n \times d_v}$$

$$\text{MH-Attn} = [\text{Attn}_1, ..., \text{Attn}_M] W^O \qquad W^O \in \mathbb{R}^{M d_v \times d_m}$$

Transformer with sparse sampling

$$X \in \mathbb{R}^{HW \times d_{model}}$$

$$Q \in \mathbb{R}^{HW \times d_{model}} \qquad V \in \mathbb{R}^{HW \times d_{model}}$$

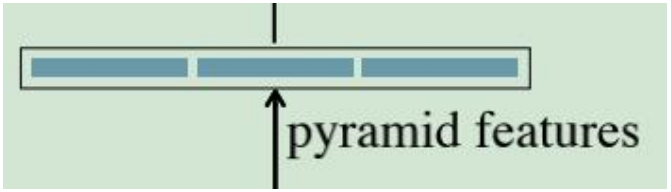
$$\text{Attn}_{mq} = \text{softmax}\left(\sum_{n=1}^{N} w_{nq}\right) v_{\langle c_q + \Delta_n \rangle}$$

O(n*n) -> O(kn)

$$A_m = \text{softmax}\left(Q W_m^Q U_m^{wts}\right) \qquad U_m^{wts} \in \mathbb{R}^{d_k \times N}$$

$$\text{Attn}_m = A_m \left(V W_m^V\right)_{\langle Q W_m^Q U_m^{pos} \rangle}$$

Pyramid transformer

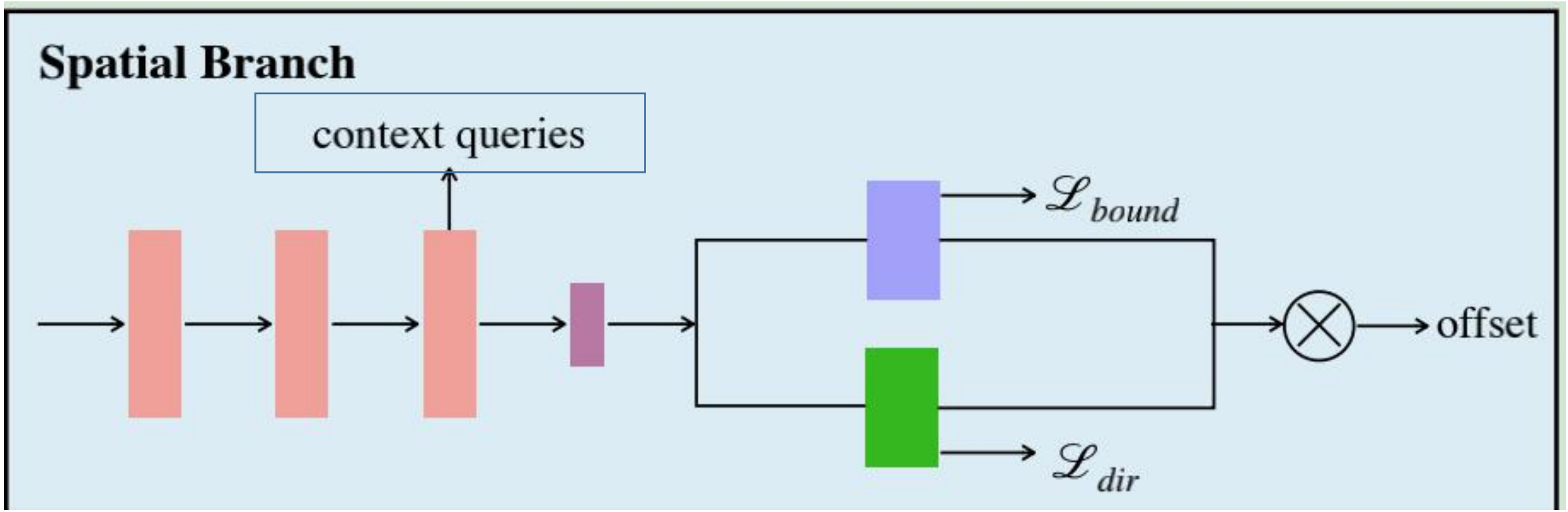


$$\{X^l\}_{l=1}^{L} \qquad X^l \in \mathbb{R}^{H_l W_l \times C}$$

$$X_{ms} \in \mathbb{R}^{L_{ms} \times d_{model}}$$

$$\text{Attn}_{mq} = \text{softmax}\left(\sum_{l=1}^{L}\sum_{n=1}^{N} w_{lnq}\right) v_{\langle c_q + \Delta_{ln} \rangle}$$

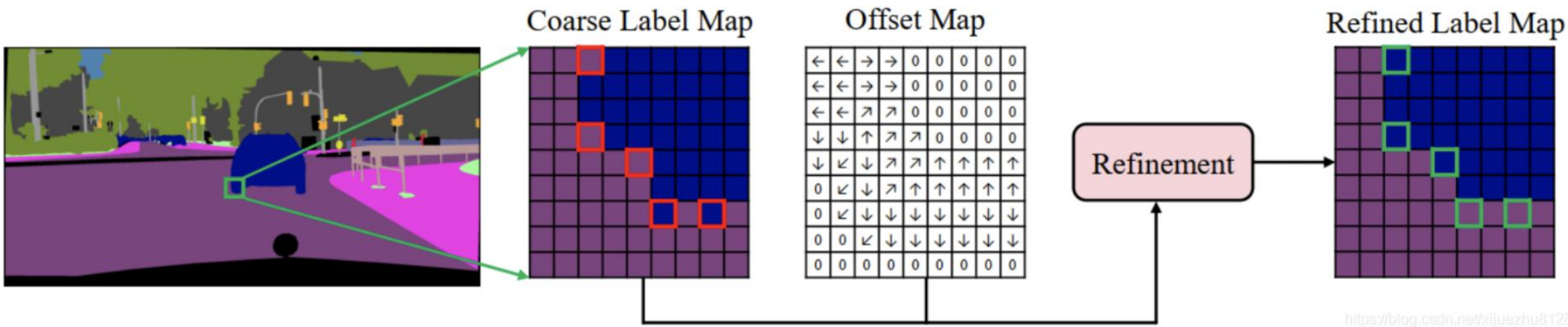
Dynamic learnable spatial branch



Boundary refinement

把边界作为一个类预测出来（语义分割）
预测边界的方向

测试时将一些边界像素标签进行重新优化



Experiments

Method	Reference	Backbone	mIoU	pixAcc
PSPNet [70]	CVPR2017	ResNet50	41.7	80.0
PSANet [71]	ECCV2018	ResNet50	42.9	80.9
UperNet [51]	ECCV2018	ResNet50	41.2	79.9
EncNet [65]	CVPR2018	ResNet50	41.1	79.7
CFNet [67]	CVPR2019	ResNet50	42.9	-
CPNet [56]	CVPR2020	ResNet50	44.5	81.4
Ours	-	ResNet-50	**46.1**	**81.7**
RefineNet [36]	CVPR2017	ResNet101	40.2	-
PSPNet [70]	CVPR2017	ResNet101	43.3	81.4
SAC [69]	ICCV2017	ResNet101	44.3	81.9
UperNet [51]	ECCV2018	ResNet101	42.7	81.0
DSSPN [35]	CVPR2018	ResNet101	43.7	81.1
PSANet [71]	ECCV2018	ResNet101	43.8	81.5
EncNet [65]	CVPR2018	ResNet101	44.7	81.7
ANL [78]	ICCV2019	ResNet101	45.2	-
CCNet [26]	ICCV2019	ResNet101	45.2	-
CFNet [67]	CVPR2019	ResNet101	44.9	-
CPNet [56]	CVPR2020	ResNet101	46.3	81.9
OCRNet [62]	ECCV2020	ResNet101	45.3	-
Efficient FCN [37]	ECCV2020	ResNet101	45.3	-
ResNeSt [66]	arXiv2020	ResNeSt200	48.4	-
SETR [72]	CVPR2021	T-large	50.2	83.5
Ours	-	DeiT	**50.5**	**83.6**

Table 1. Quantitative evaluations on the ADE20K validation set.

Experiments

Method	Backbone	mIoU
PSPNet [70]	ResNet101	78.5
DeepLabv3 [11] (MS)	ResNet101	79.3
PointRend [29]	ResNet101	78.3
OCRNet [62]	ResNet101	80.6
Multiscale DEQ [2] (MS)	MDEQ	80.3
CCNet [26]	ResNet101	80.2
GCNet [6]	ResNet101	78.1
Axial-DeepLab-XL [49] (MS)	Axial-ResNet-XL	81.1
Axial-DeepLab-L [49] (MS)	Axial-ResNet-L	81.5
SETR [72] (MS)	T-large	82.2
Ours (80k, MS)	ResNet50	79.8
Ours (80k, MS)	DeiT	**82.9**

Table 2. Quantitative evaluations on the Cityscapes validation set (training iterations: 80k, MS: Multi-scale inference).

Method	Reference	Backbone	mIoU
FCN-8S [41]	CVPR2015	VGG16	37.8
BoxSup [17]	ICCV2015	VGG16	40.5
RefineNet [36]	CVPR2017	ResNet152	47.3
PSPNet [70]	CVPR2017	ResNet101	47.8
EncNet [65]	CVPR2018	ResNet101	51.7
DANet [21]	CVPR2019	ResNet101	52.6
ANL [78]	ICCV2019	ResNet101	52.8
CPNet [56]	CVPR2020	ResNet101	53.9
OCRNet [62]	ECCV2020	ResNet101	54.8
Efficient FCN [37]	ECCV2020	ResNet101	55.3
SETR [72]	CVPR2021	T-large	**55.8**
Ours	-	ResNet50	49.5
Ours	-	DeiT	55.2

Table 4. Quantitative evaluations on the PASCAL-Context validation set.

Ablation studies

Variants	pixAcc	mIoU
baseline	80.3	42.3
+ pyramid features	81.1	45.0
+ spatial path	81.7	46.1
+ stronger backbone	**83.6**	**50.5**

Table 5. Evaluation of different components on ADE20K validation set.

Feature scales (L)	Sampling points (N)	pAcc	mIoU
1	4	77.5	37.5
	16	80.3	42.5
	64	80.5	42.9
3	4	80.6	44.1
	16	**81.7**	**46.1**
	64	80.4	45.6

Table 6. Ablation studies on different sampling points (N) and different feature scales (L) on ADE20K dataset with a ResNet50 backbone. Note that, boundary refinement is not applied here.

Ablation studies

Method	Backbone	Mem Cost	mIoU
Sparse Transformer [14]	ResNet50	18.6G	40.3
Longformer [3]	ResNet50	11.3G	39.4
Reformer [30]	ResNet50	15.5G	38.2
CCNet [26]	ResNet50	9.8G	43.1
ANL [78]	ResNet50	2.0G	42.6
SETR [72]	T-large	30.0G	50.2
Ours	ResNet50	7.0G	46.1
Ours	DeiT	8.5G	**50.5**

Table 8. Ablation studies on the memory efficiency of UN-EPT. We report results on ADE20K validation set.

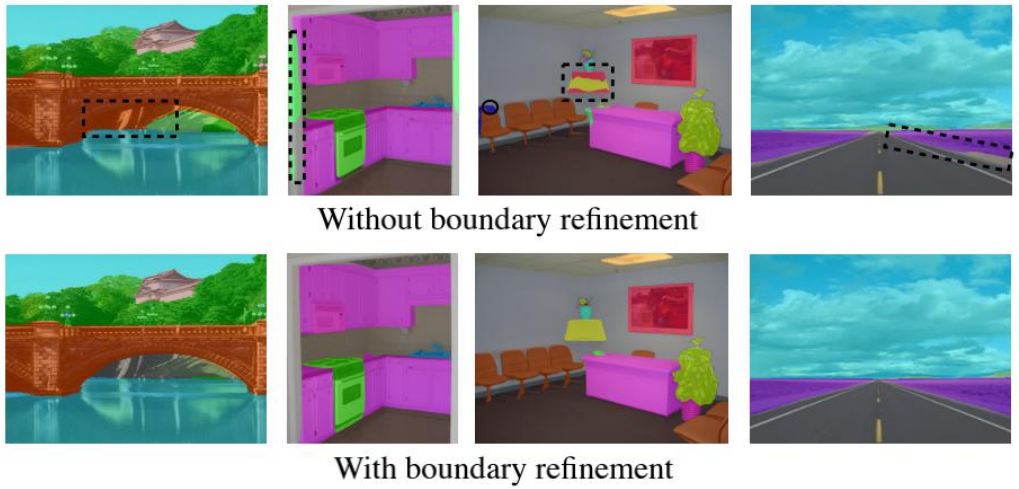


Figure 4. Visualization of segmentation results with/without boundary refinement. Examples are taken from ADE20K validation set.

谢谢大家!