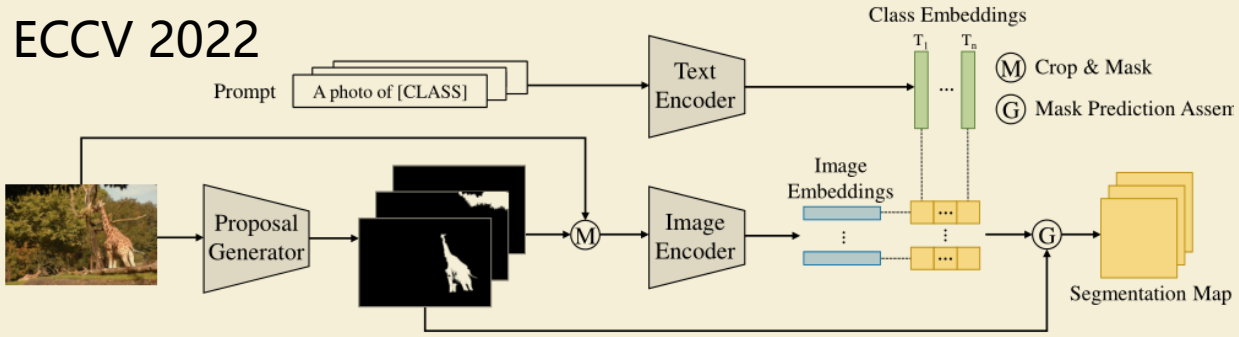
Open-Vocabulary Segmentation

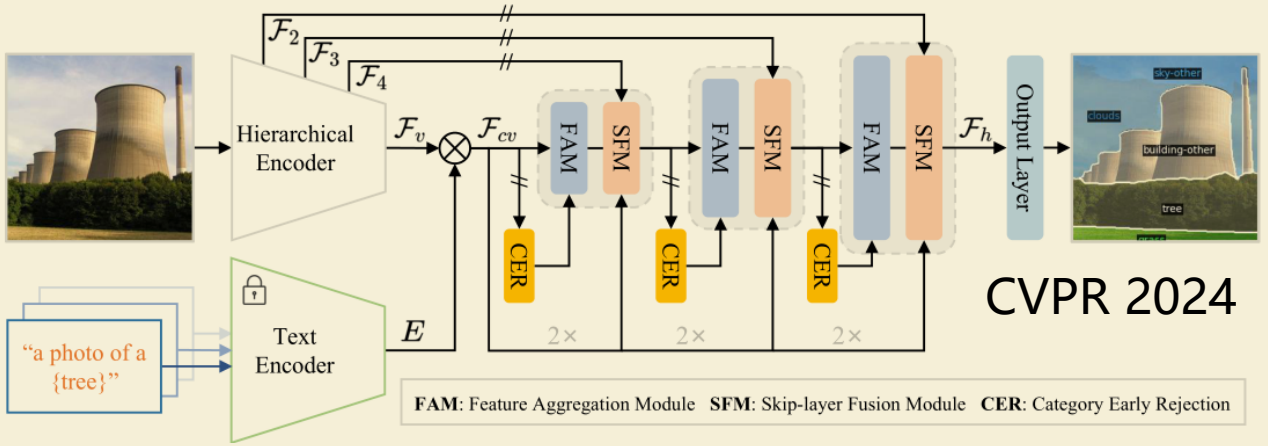
常世杰

2023.3.31

Fully-Supervised OV SS

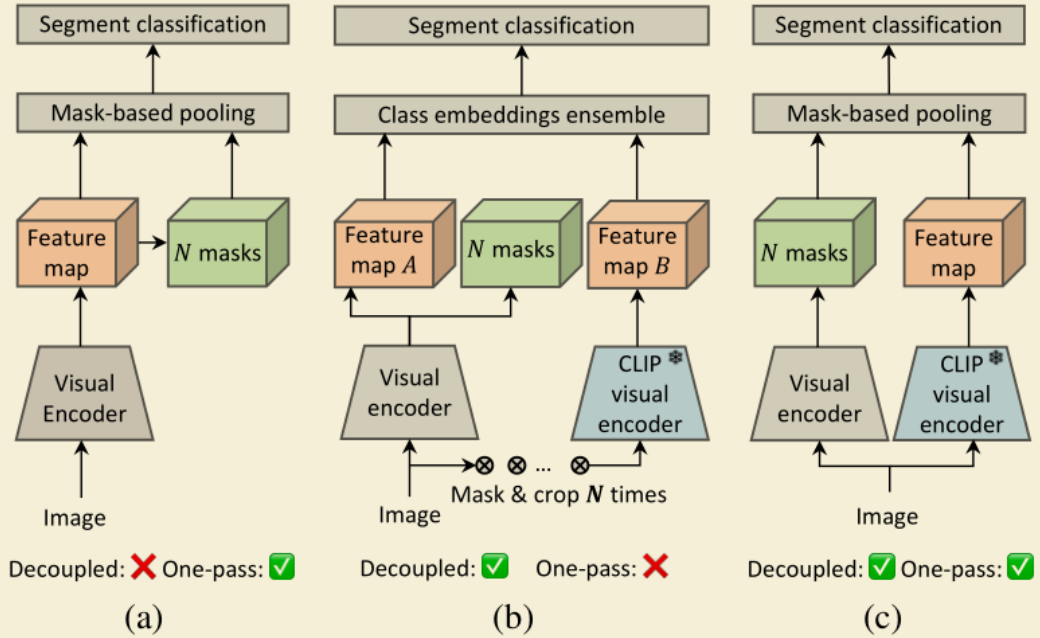
ECCV 2022





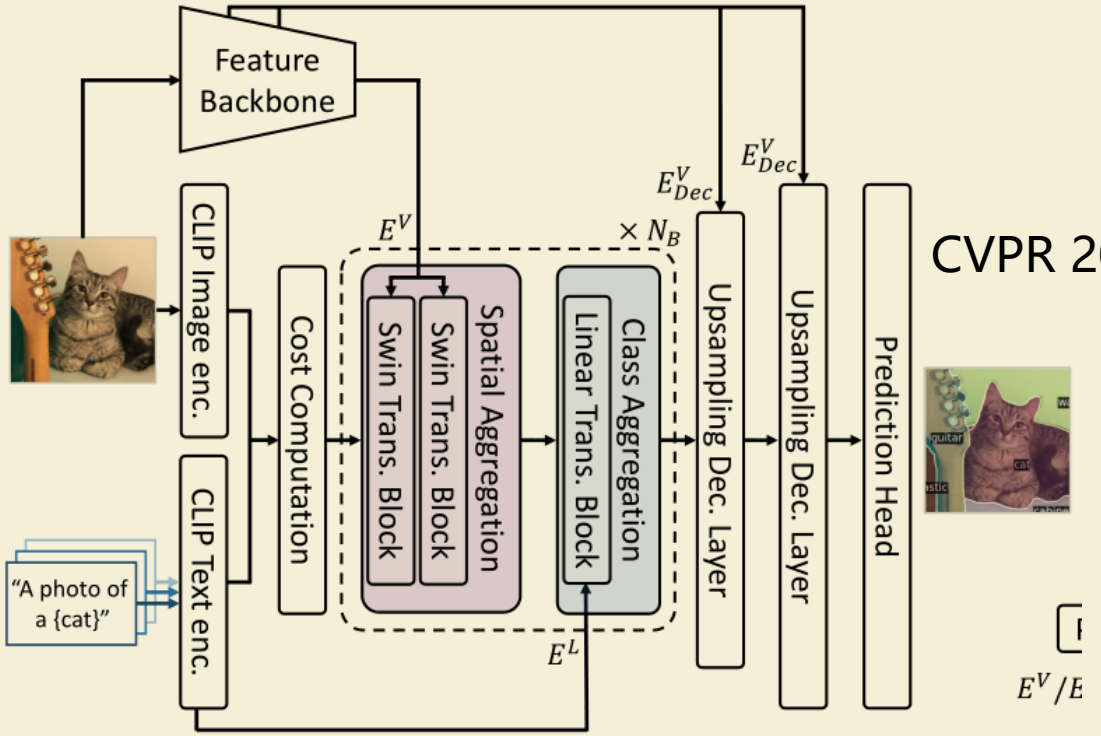
CVPR 2024

Two-stage OVSeg



ICCV 2023

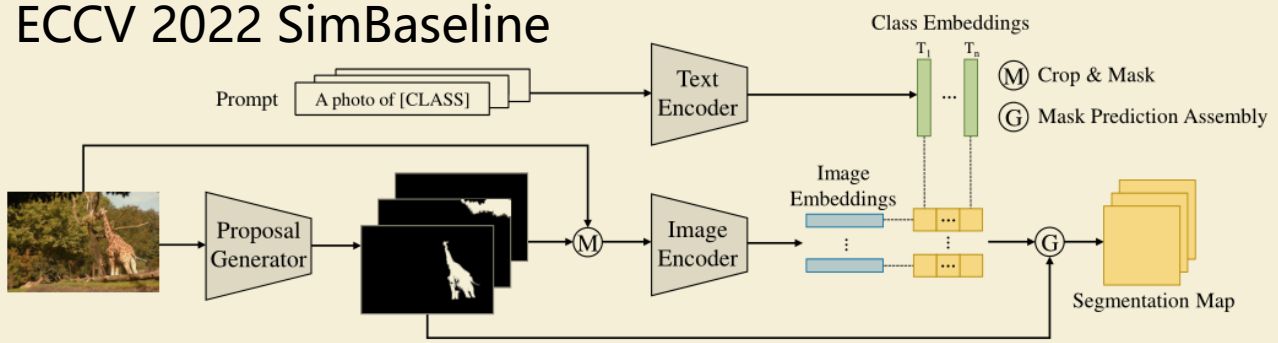
One-stage OVSeg



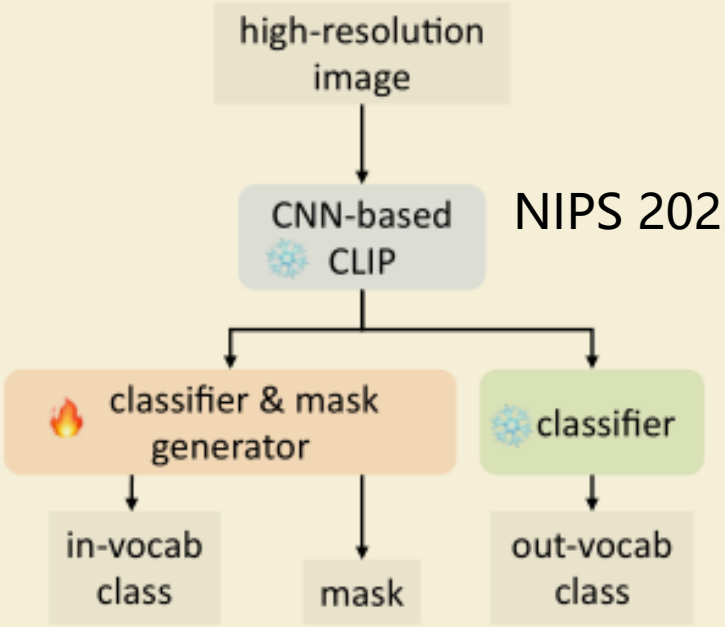
CVPR 2024

Two-stage OVSeg

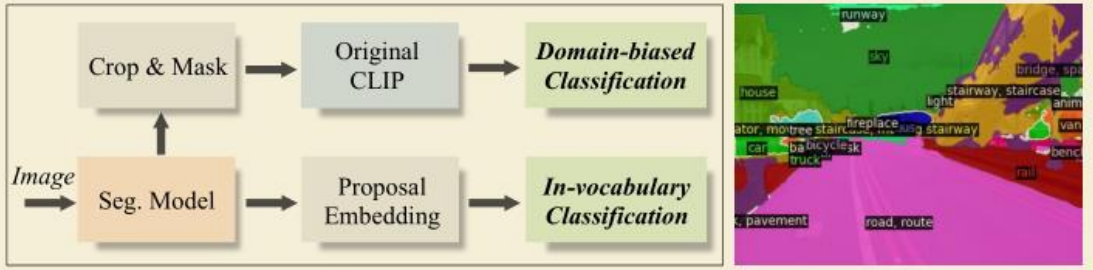
ECCV 2022 SimBaseline



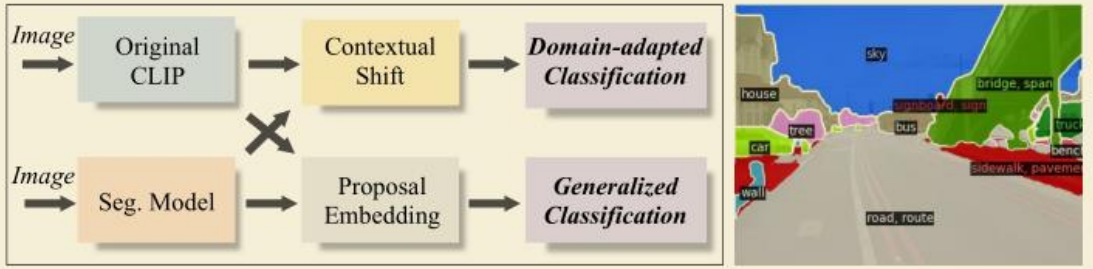
FC-CLIP (ours):



NIPS 2023 FC-CLIP



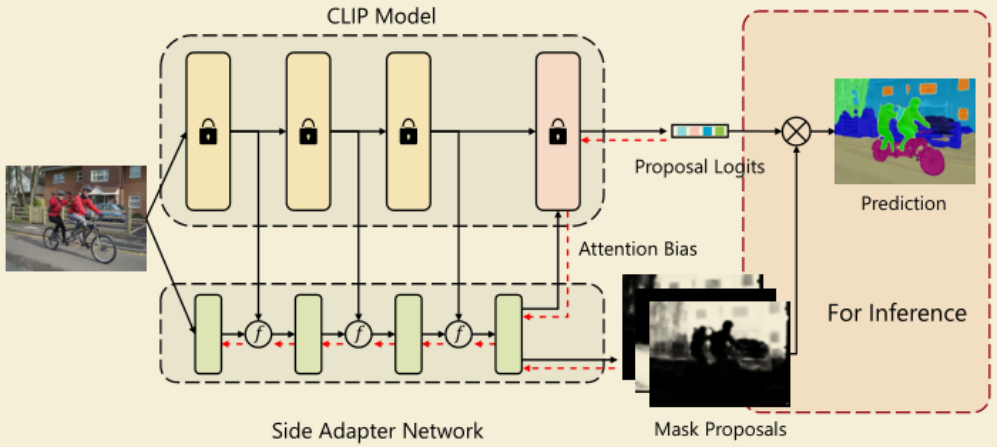
(a) existing two-stage methods



(b) our semantic-assisted calibration network

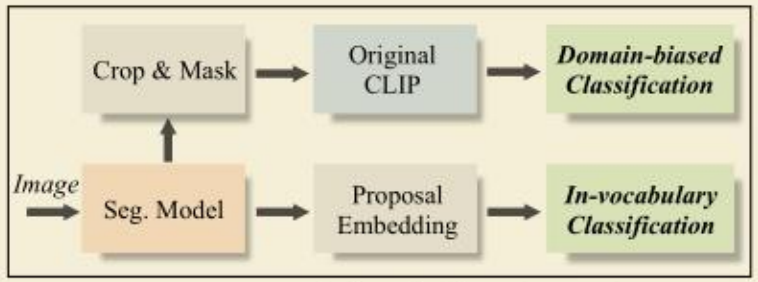
CVPR 2024 SCAN

Mask proposal (mask2former) + CLIP Classification

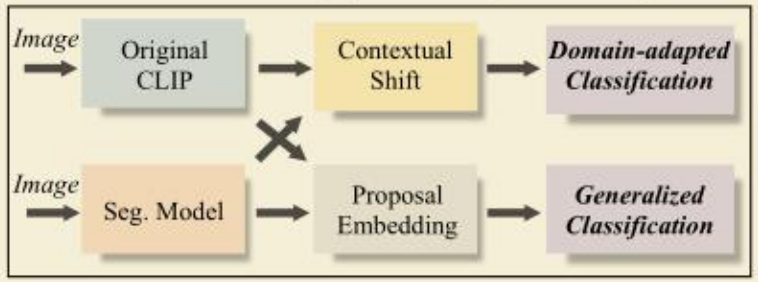


CVPR 2023 SAN

Open-Vocabulary Segmentation with Semantic-Assisted Calibration



(a) existing two-stage methods



natural image

masked image

Ori. CLIP Pred: 0.41 van | 0.30 truck | 0.17 car ..

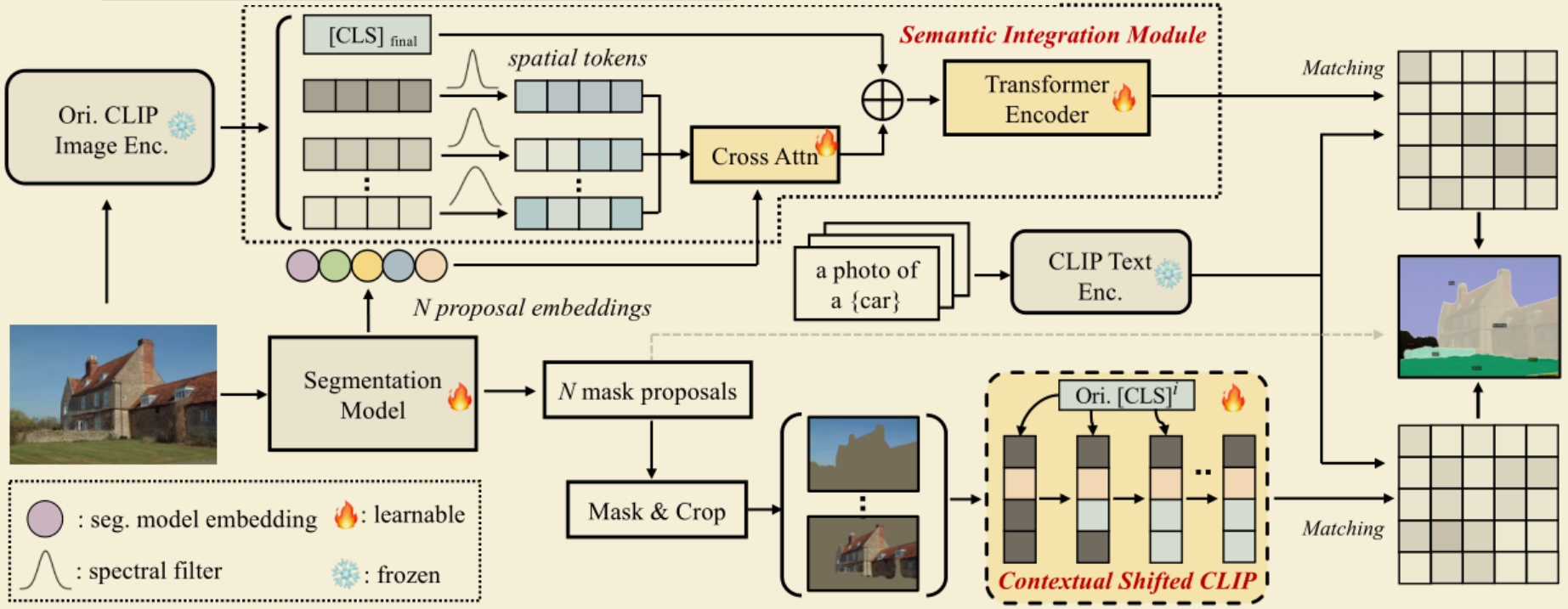
CS. CLIP Pred: 0.71 road | 0.07 dirt | 0.03 sidewalk ..

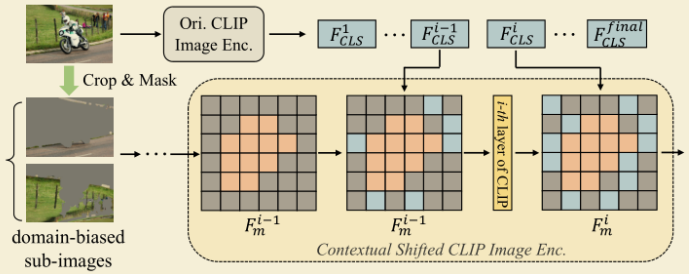
natural image

masked image

Ori. CLIP Pred: 0.999 plane | 7e-4 sky | 6e-5 runway ..

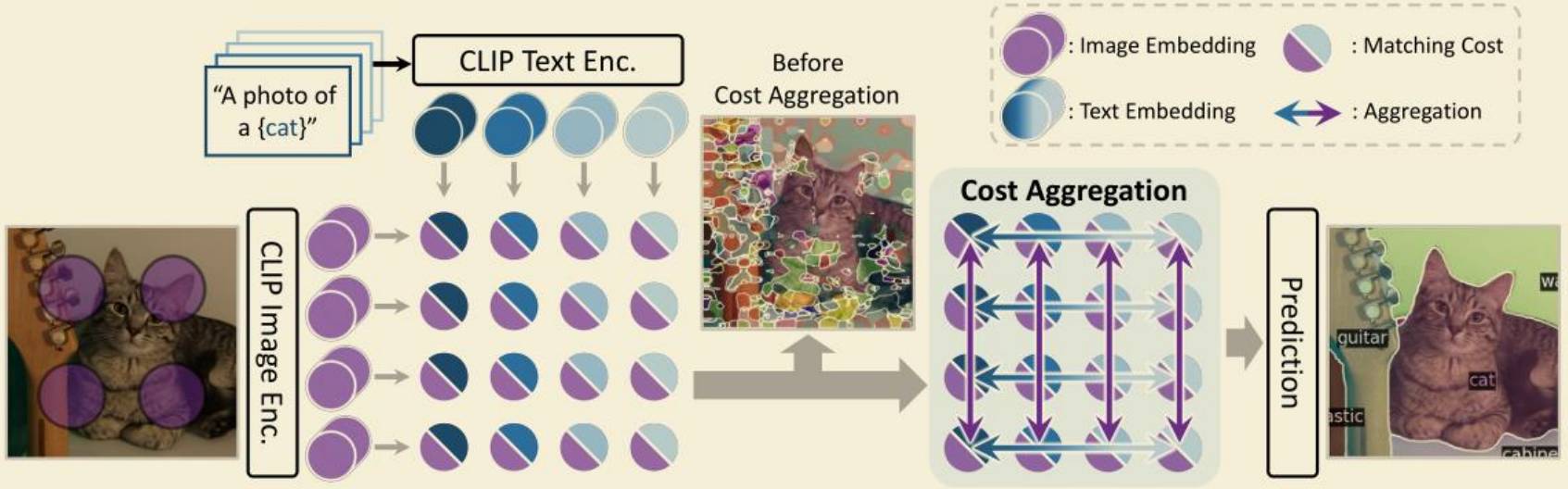
CS. CLIP Pred: 0.989 sky | 5e-3 plane | 1e-3 canopy ..



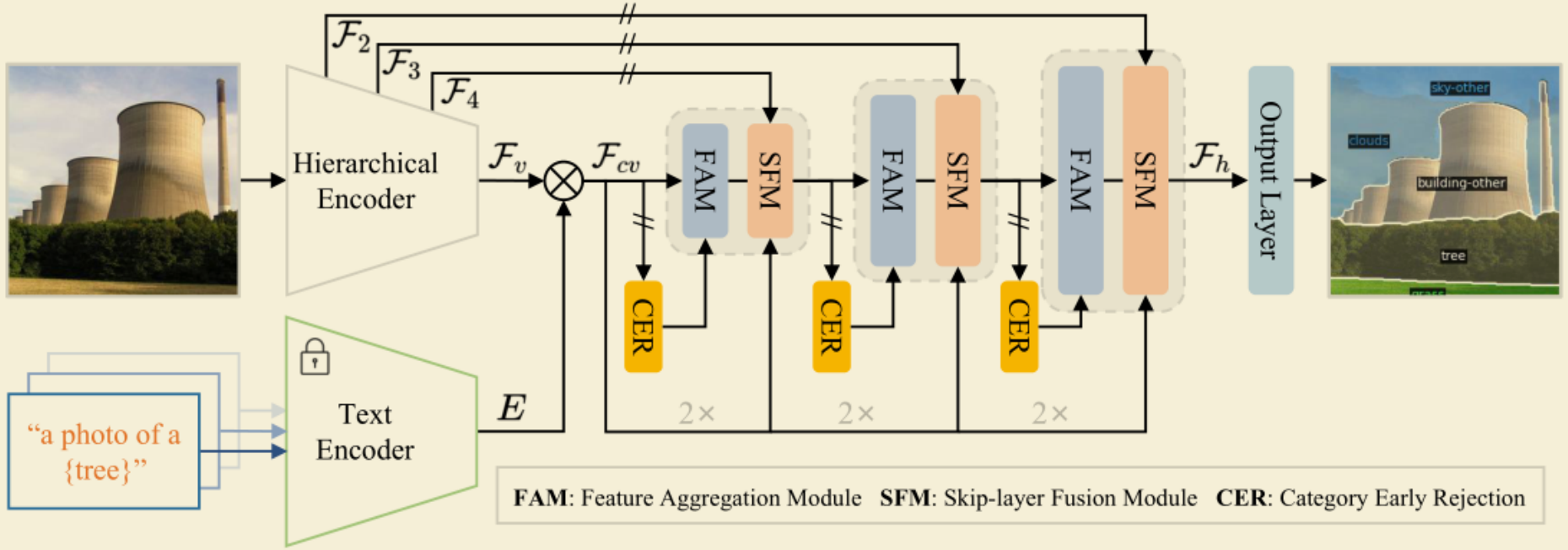


One-stage OVSeg

CVPR 2024
CAT-Seg

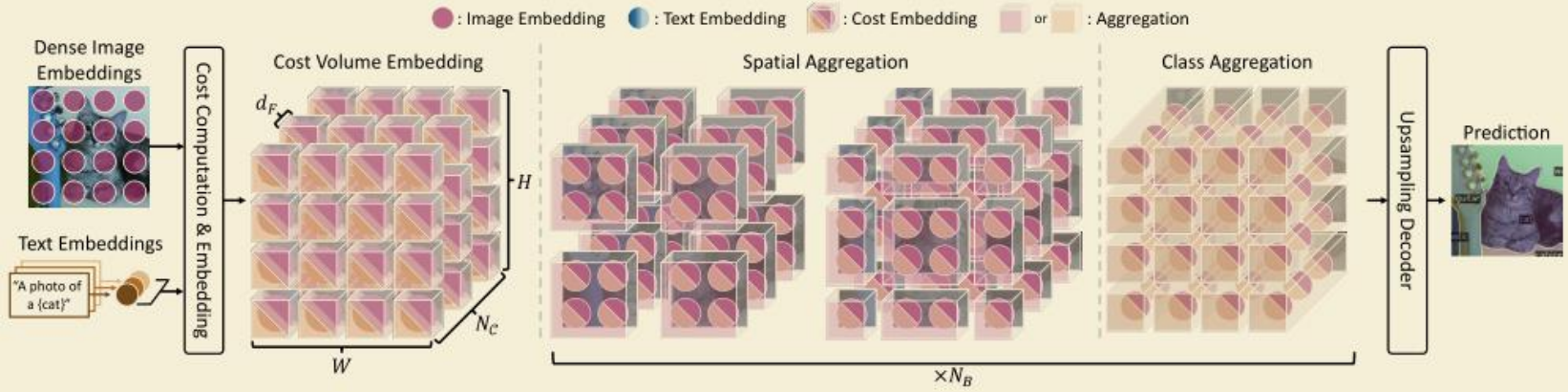


CVPR 2024 SED

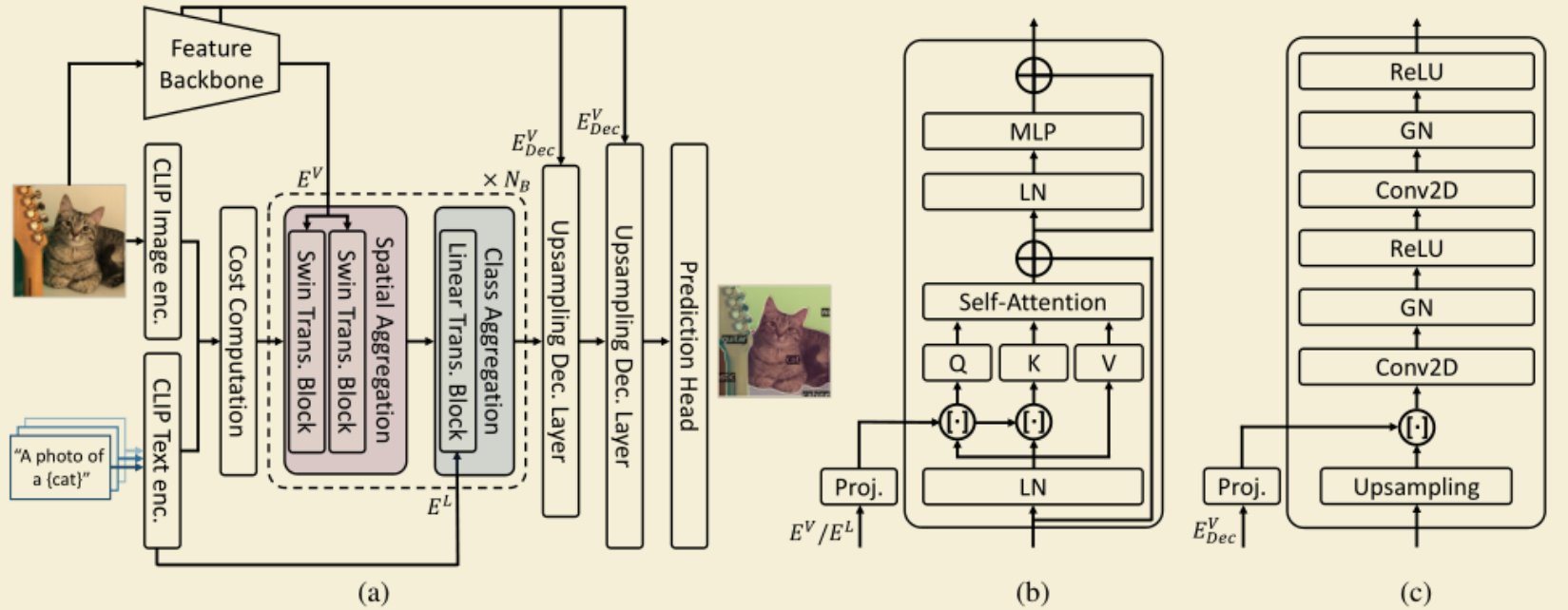


CAT-Seg🐱: Cost Aggregation for Open-Vocabulary Semantic Segmentation

Seokju Cho*,1 Heeseong Shin*,1 Sunghwan Hong[1] Seungjun An[1] Seungjun Lee[1],
Anurag Arnab[2] Paul Hongsuck Seo[2] Seungryong Kim[1]
[1]Korea University [2]Google Research

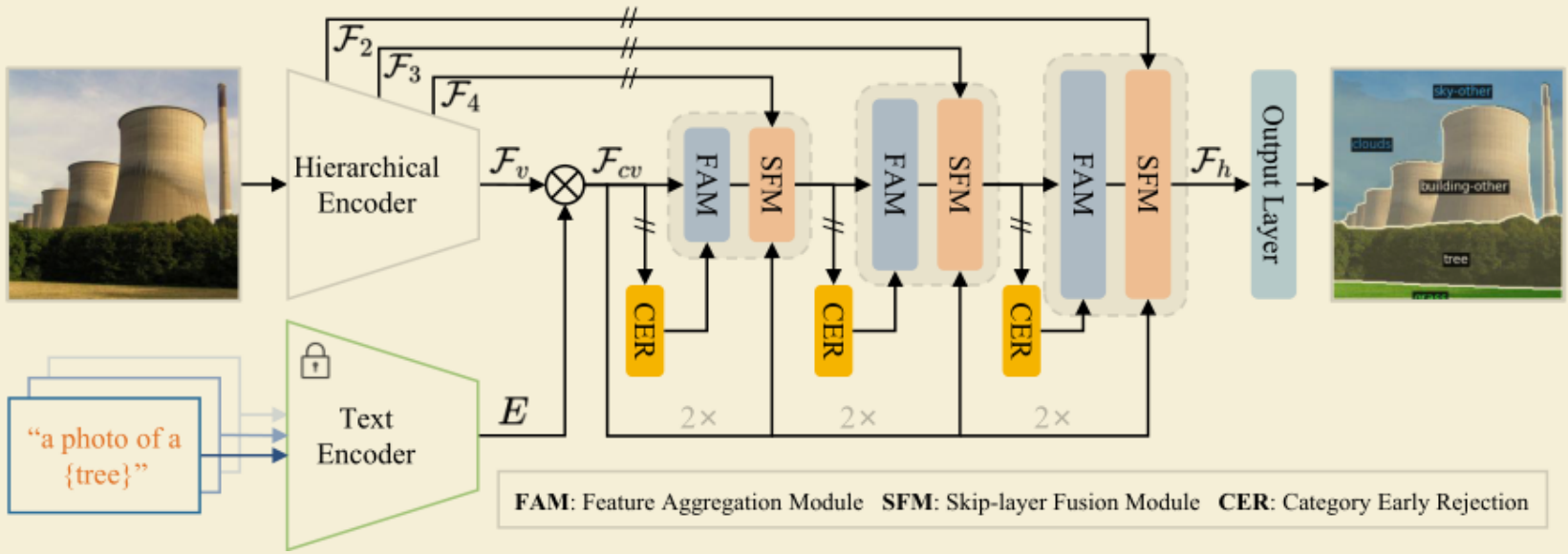


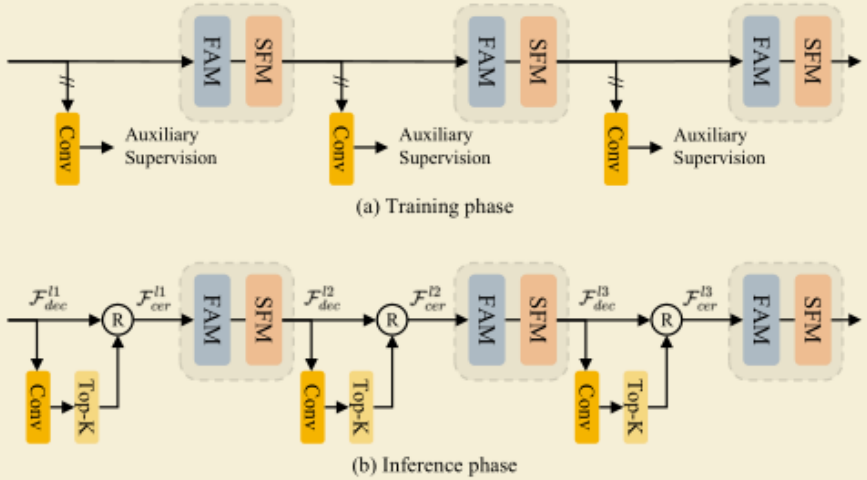
N, H, W



SED: A Simple Encoder-Decoder for Open-Vocabulary Semantic Segmentation

Bin Xie[1], Jiale Cao[1], Jin Xie[2], Fahad Shahbaz Khan[3], and Yanwei Pang[1,4]
[1]Tianjin University [2]Chongqing University





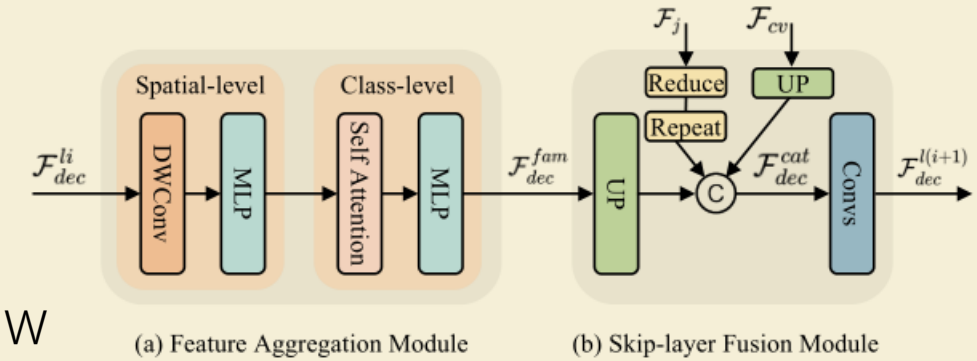
双阶段法用两个独立的网络预测mask和类别—速度慢

现有单阶段法用ViT作为视觉backbone—局部表示差

利用层次化结构带来更多局部信息

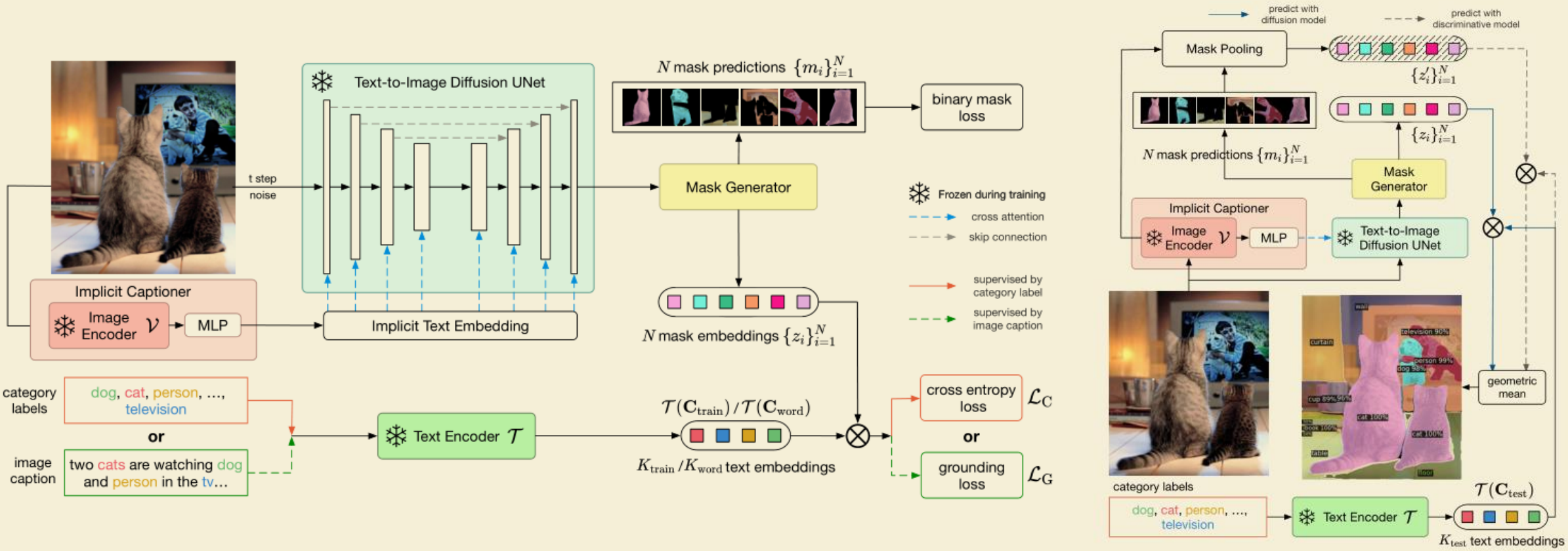
NPC
CHW
NDHW N1HW

提出CER，提前过滤图片中不存在的类别，提高推理速度

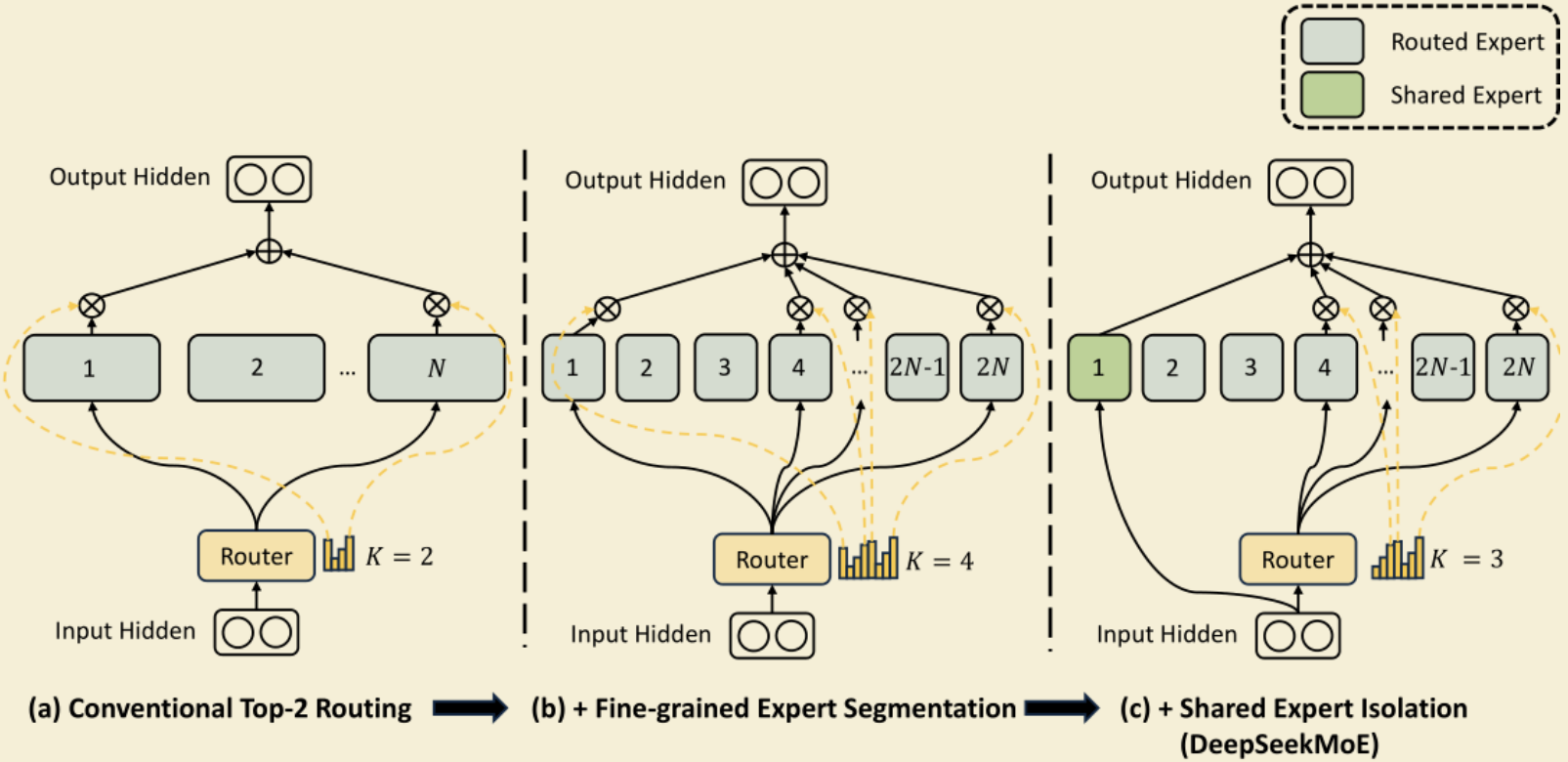


Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

Jiarui Xu[1*] Sifei Liu[2†] Arash Vahdat[2†] Wonmin Byeon[2]
Xiaolong Wang[1] Shalini De Mello[2]
[1]UC San Diego [2]NVIDIA

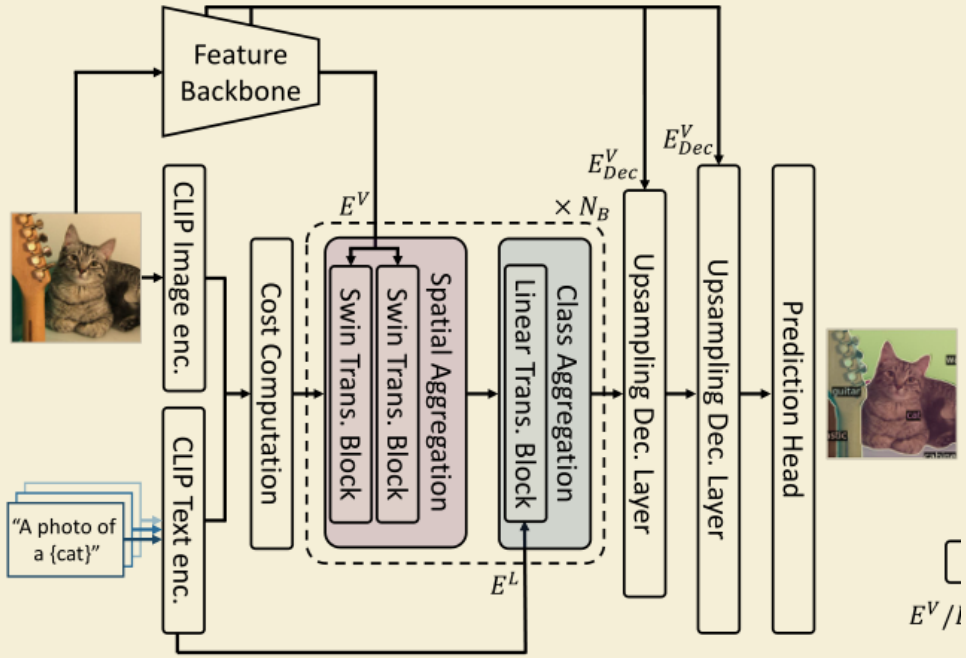


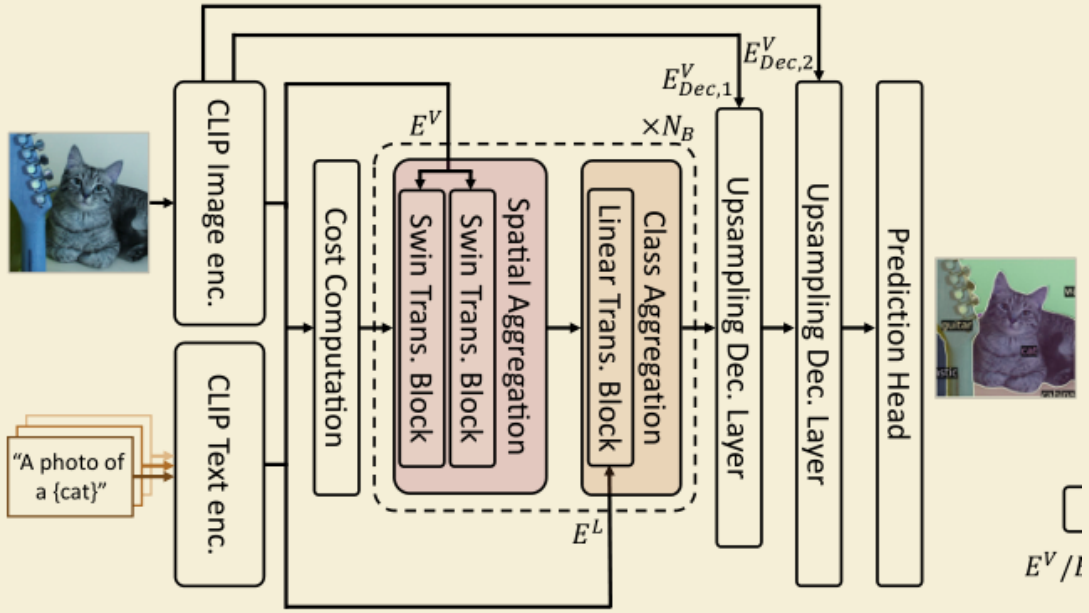
DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models



(a) Conventional Top-2 Routing → **(b) + Fine-grained Expert Segmentation** → **(c) + Shared Expert Isolation (DeepSeekMoE)**

	A-847	PC-459	A-150	PC-59	PAC-20
SED-B	11.4	18.6	31.6	57.3	94.4
moe	10.9	17.7	31.5	57.5	94.1
deepseek	10.8	18.1	31.5	57.1	93.6





	A-847	PC-459	A-150	PC-59	PAC-20
CAT-v1-B	8.4	16.6	27.2	57.5	93.7
SED-B	11.4	18.6	31.6	57.3	94.4
CAT-v2-B	12.0	19.0	31.8	57.5	94.6
SED-L	13.9	22.6	35.2	60.6	96.1
CAT-v2-L	16.0	23.8	37.9	63.3	97.0

特征图基于cost volume，导致特征图大小受类别数影响(B, C, H, W) -> (B, T, C, H, W) solov2

预测逻辑：为每个类别T预测一个二值mask

训练集和测试集类别域偏移太大，测试集很多类别IoU=0

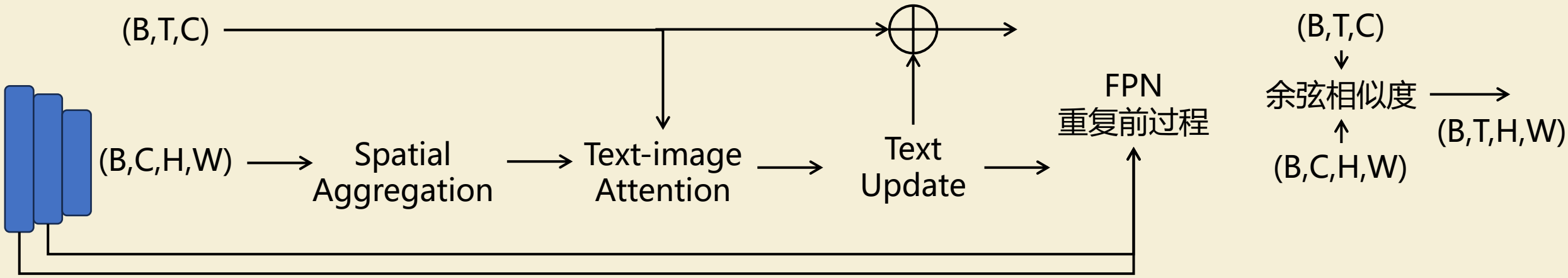
CAT-Seg-v1 v.s. SED v.s. CAT-Seg-v2

CAT-Seg-v1
CLIP-ViT + feature backbone
Fine-tune visual encoder attention
Train 384
Test sliding inference 384/640
Imagenet templates

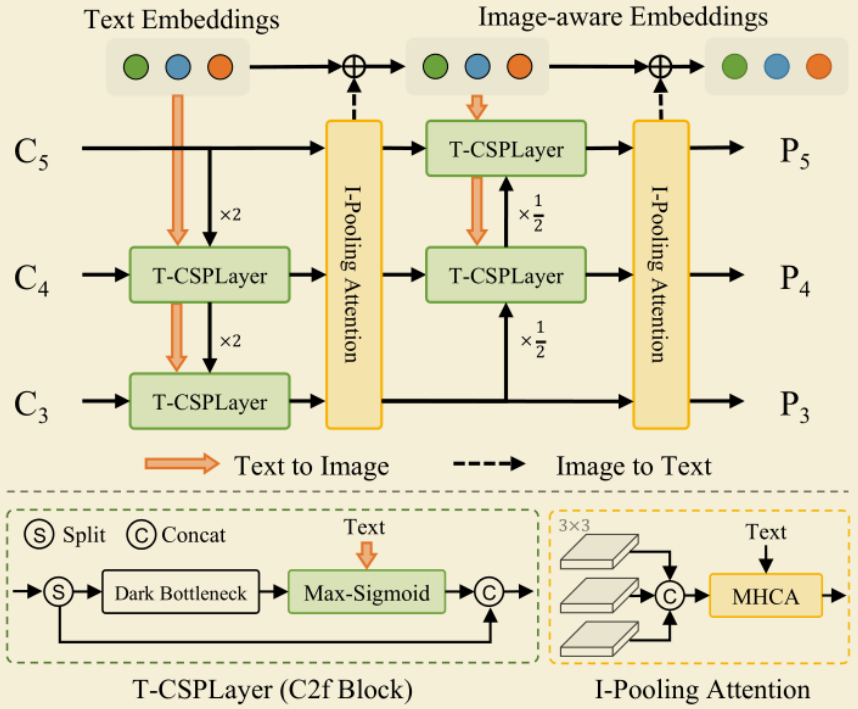
SED
CLIP-ConvNext
Fine-tune full visual encoder
Train 768
Test 640
Imagenet templates

CAT-Seg-v2
CLIP-ViT
Fine-tune q/v visual&textual encoder
Train 384
Test sliding inference 384/640
Single template

ViT-Base 2d8h



	A-847	PC-459	A-150	PC-59	PAC-20
SED-B	11.4	18.6	31.6	57.3	94.4
	3.1	9.0	17.3	49.3	89.1

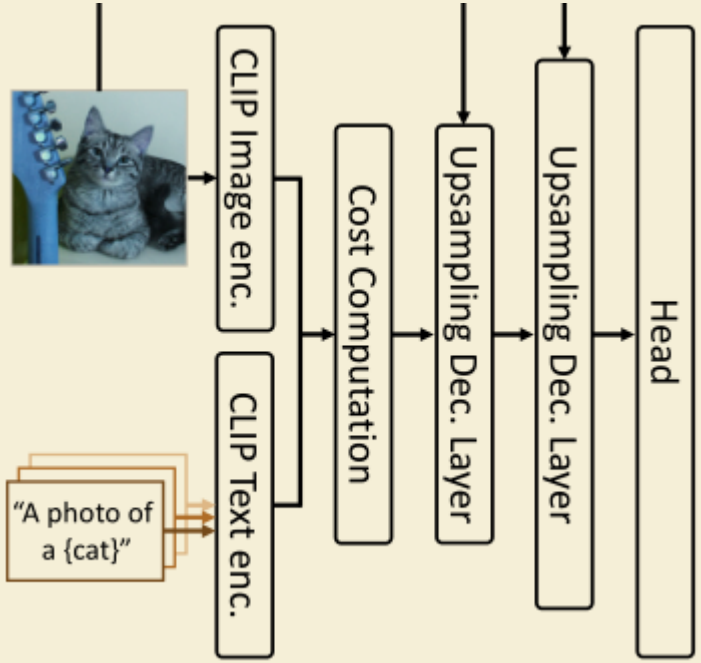


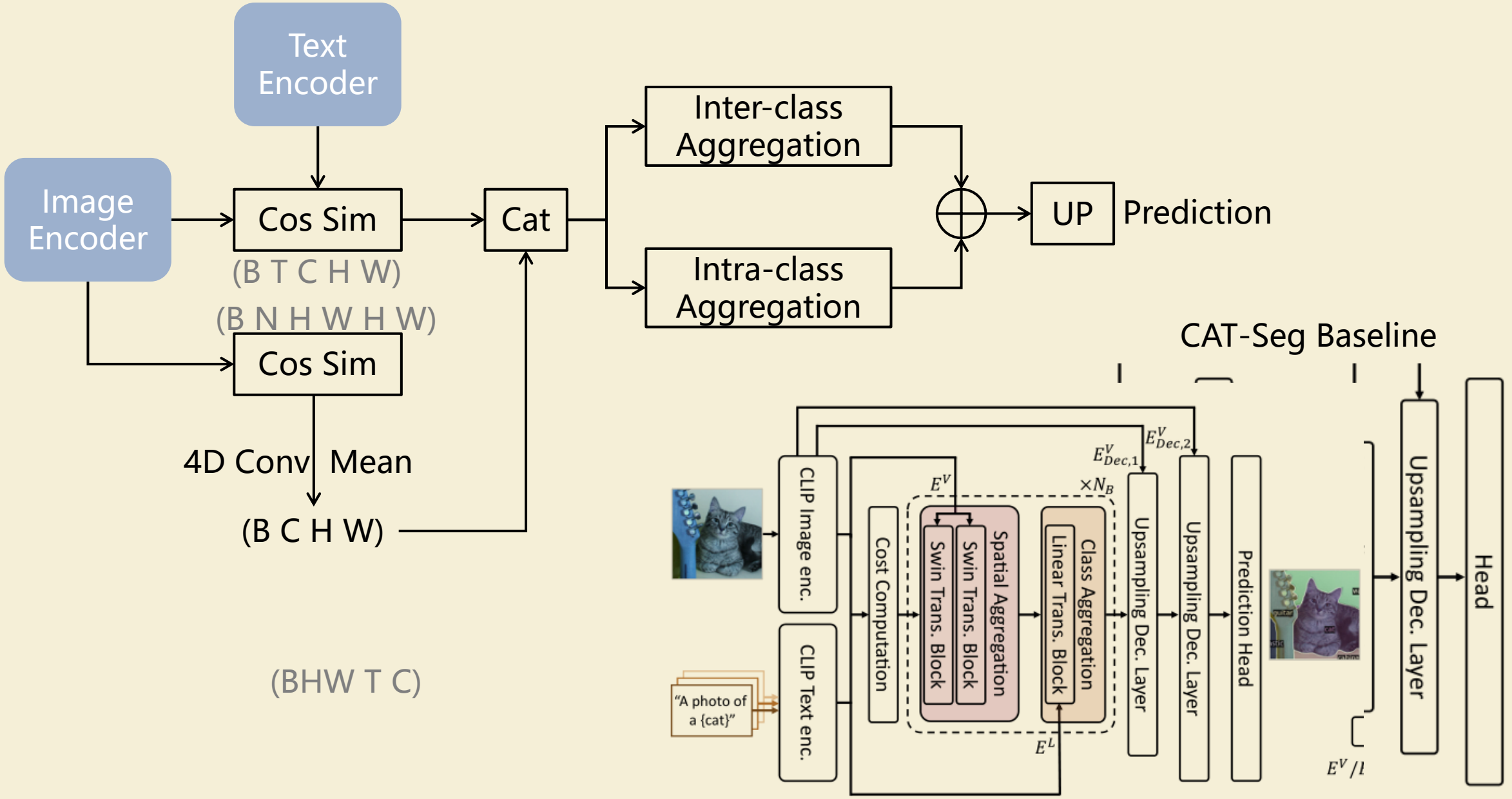
	A-847 8.00	PC-459 20	A-150 5.31	PC-59 9.01	PAC-20 2.0
CAT-B	12.1	19.1	31.6	57.4	94.7
baseline	10.0	16.6	27.0	52.6	93.8
baseline2	10.8	17.5	28.6	54.0	94.2
	11.1	17.8	30.7	56.0	94.4
	11.6	18.6	31.1	56.8	94.8
dino	10.1	16.5	27.1	52.5	93.9
deformcl	11.6	18.9	31.5	57.2	94.8

	A-847 8.00	PC-459 20	A-150 5.31	PC-59 9.01	PAC-20 2.0
baseline2	10.8	17.5	28.6	54.0	94.2
+aug	10.9	17.7	28.6	54.4	94.6
+ms	10.8	17.5	28.4	53.9	94.5
dcn	11.4	18.3	31.1	55.7	94.8
+cls	12.0	19.5	31.9	58.2	94.4
Oracle	28.7	43.2	51.1	75.3	97.7

	Components	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
(I)	Feature Agg.	5.6	12.8	23.6	58.1	96.3	77.7
(II)	Cost Agg.	14.7	23.2	35.3	60.3	96.7	78.9
(III)	(II) + Spatial agg.	14.9	23.1	35.9	60.3	96.7	79.5
(IV)	(II) + Class agg.	14.7	21.5	36.6	60.6	95.5	80.5
(V)	(II) + Spatial and Class agg.	15.5	23.2	37.0	62.3	96.7	81.3
(VI)	(V) + Embedding guidance	**16.0**	**23.8**	**37.9**	**63.3**	**97.0**	**82.5**

-	-	-	-	52.3	-
8.1	11.5	26.4	44.8	-	70.2
9.0	12.4	29.6	55.7	94.5	-
12.4	15.7	32.1	57.7	94.6	-
11.1	14.5	29.9	57.3	-	-
16.0	**23.8**	**37.9**	**63.3**	**97.0**	**82.5**
(+3.6)	(+8.1)	(+5.8)	(+5.6)	(+2.4)	(+12.3)




Text Encoder
Image Encoder
Cos Sim
(B T C H W)
(B N H W H W)
Cos Sim
4D Conv
Mean
(B C H W)
(BHW T C)
Cat
Inter-class Aggregation
Intra-class Aggregation
UP
Prediction
CAT-Seg Baseline
CLIP Image enc.
CLIP Text enc.
"A photo of a {cat}"
Cost Computation
E^V
$\times N_B$
Swin Trans. Block
Swin Trans. Block
Spatial Aggregation
Linear Trans. Block
Class Aggregation
E^L
$E^V_{Dec,1}$
$E^V_{Dec,2}$
Upsampling Dec. Layer
Upsampling Dec. Layer
Prediction Head
$E^V/1$
Upsampling Dec. Layer
Head

	A-847 8.00	PC-459 20	A-150 5.31	PC-59 9.01	PAC-20 2.0
CAT-B	12.1	19.1	31.6	57.4	94.7
baseline	10.0	16.6	27.0	52.6	93.8
baseline2	10.8	17.5	28.6	54.0	94.2
i-i corr	11.1	17.8	30.7	56.0	94.4
i-i corr2	11.6	18.6	31.1	56.8	94.8
dino	10.1	16.5	27.1	52.5	93.9
maskMHA	11.6	18.9	31.5	57.2	94.8
dcn	11.4	18.3	31.1	55.7	94.8

去掉spatial和class aggregation和decoder中的浅层特征

去掉spatial和class aggregation

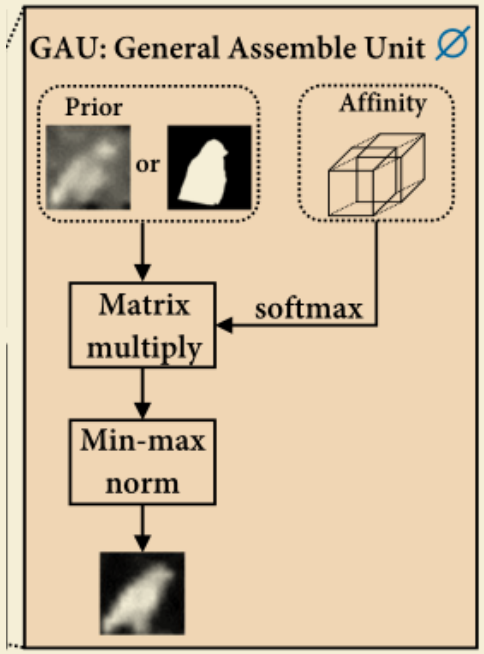
加入i-i correlation

在baseline2的基础上将浅层特征替换为dino对应层特征

Class aggregation时使用mask MHA，仅和相近的K个类别做MHA
K初步设置为16，感觉应该增大K试一下
降点原因可能是原spatial agg都使用了大核 CAT-Seg 12，SED 9
初始DCNv3核设置为3

- 获取更好的初始化cost volume
- 更好的cost aggregation方式

对spatial和class都使用deformable操作：
Spatial，改一下DCNv3中生成offset和mask的过程
Class，使用mask MHA仅和相近的K个类别做MHA



	A-847 8.00	PC-459 20	A-150 5.31	PC-59 9.01	PAC-20 2.0
CAT-B	12.1	19.1	31.6	57.4	94.7
oracle	28.7	43.2	51.1	75.3	97.9
Cls_head	12.2	19.2	31.9	58.1	94.7
APL	12.1	19.1	32.0	58.2	94.9
MLDecoder	12.1	19.5	31.5	58.0	95.0
+RAM	11.3	19.1	31.6	57.4	94.8

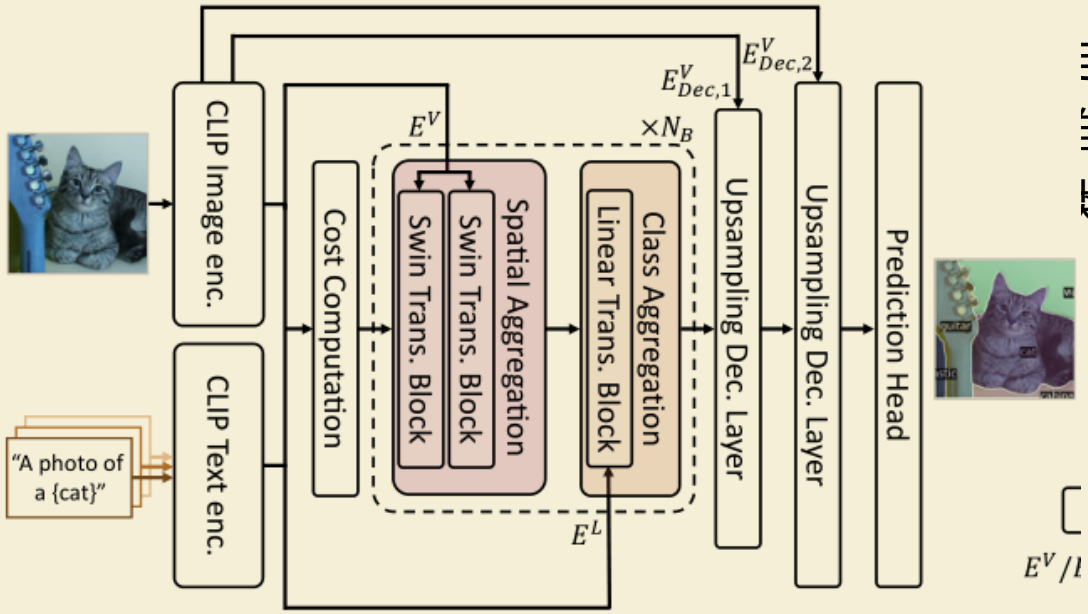
给定图片中存在的类别

Transformer cross-attention + BCE Loss

Transformer cross-attention + AP Loss

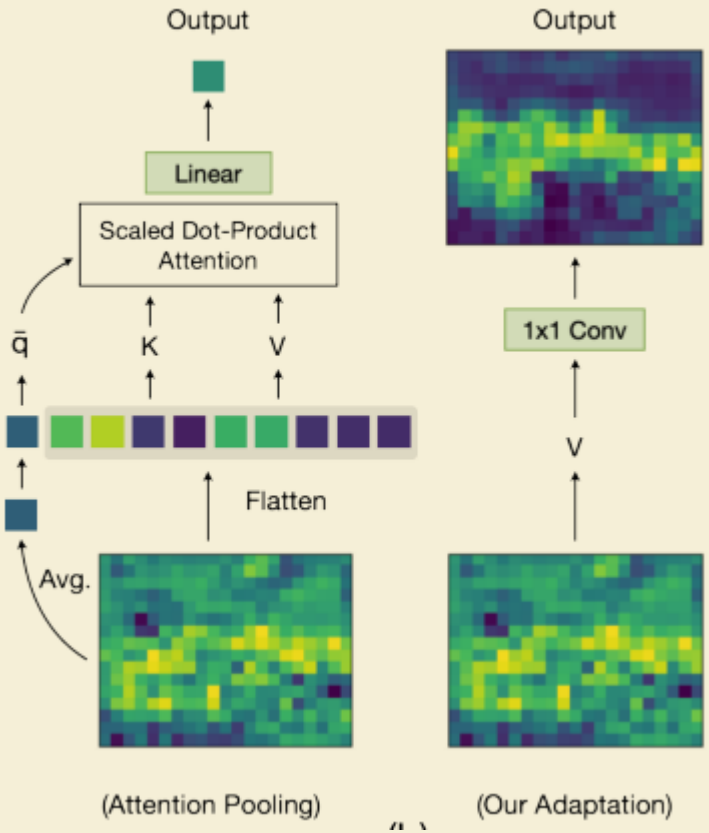
MLDecoder+ AP Loss

使用RAM进行分类直接测试性能

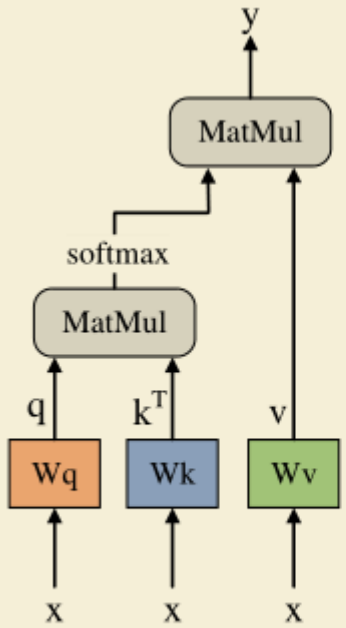


与分类联合训练后，性能会有一点提升，但过用分类头生理类别则效果不好。分类头不能做到囊括所有存在的类别

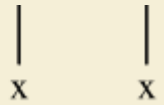
调整CLIP以生成更好的cost volume



ECCV 2022 MaskCLIP



original self-attn



corr. self-attn

$$\boldsymbol{Attn} = \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_q\boldsymbol{W}_q^T\boldsymbol{X}^T/\tau\right) + \text{Softmax}\left(\boldsymbol{X}\boldsymbol{W}_k\boldsymbol{W}_k^T\boldsymbol{X}^T/\tau\right),$$

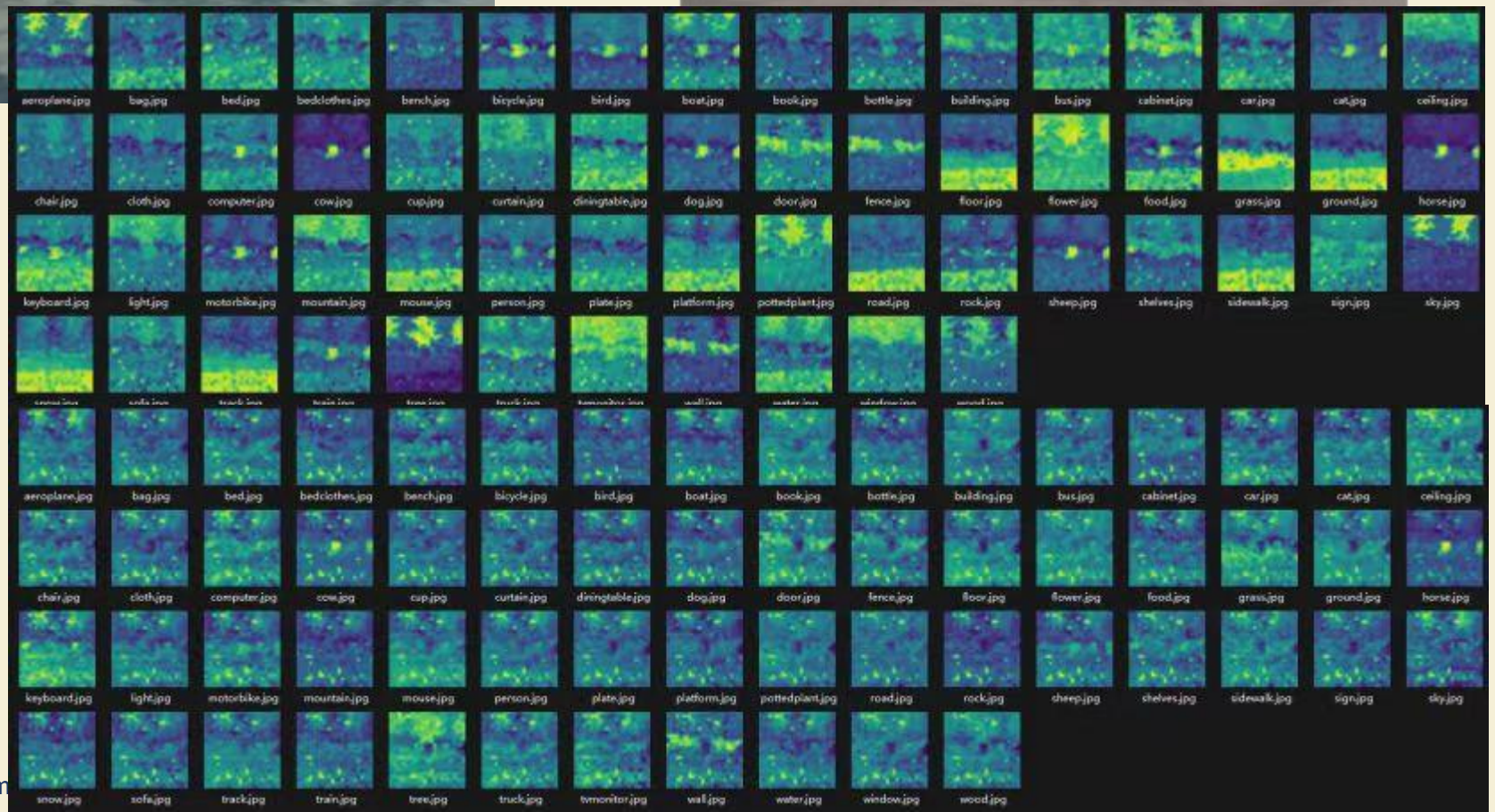
SCLIP

Method		Fair	V21	PC60	C-Obj	V20	City	PC59	ADE	C-Stf	Avg
CLIP [27]	ICML'21	✓	18.6	7.8	6.5	49.1	6.7	11.2	3.2	7.2	13.8
MaskCLIP [46]	ECCV'22	✓	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
ReCo [28]	NeurIPS'22	✗	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
GroupViT [36]	CVPR'22	✗	52.3	18.7	27.5	79.7	18.5	23.4	10.4	15.3	30.7
TCL [7]	CVPR'23	✗	55.0	30.4	31.6	83.2	24.3	33.9	17.1	22.4	37.2
CLIP Surgery [18]	Arxiv'23	✓	41.2	30.5	-	-	31.4	-	12.9	21.9	-
SCLIP [32]	Arxiv'23	✓	61.7	31.5	32.1	**83.5**	34.1	36.1	17.8	23.9	40.1
GEM [4]	CVPR'24	✓	46.2	-	-	-	-	32.6	15.7	-	-
CLIP-DIY [35]	WACV'24	✗	59.0	-	30.4	-	-	-	-	-	-
FOSSIL [3]	WACV'24	✗	-	-	-	-	23.2	35.8	18.8	24.8	-
NACLIP [PAMR]	Ours	✓	58.8	32.2	33.2	77.1	35.5	35.2	17.4	22.9	39.0
NACLIP	Ours	✓	**62.4**	**35.0**	**36.2**	80.6	**38.3**	**38.4**	**19.1**	**25.2**	**41.9**

NACLIP


sky
tree
horse
grass
und


sky
tree
building
wood
cow
grass


aeroplane.jpg
bag.jpg
bed.jpg
bedclothes.jpg
bench.jpg
bicycle.jpg
bird.jpg
boat.jpg
book.jpg
bottle.jpg
building.jpg
bus.jpg
cabinet.jpg
car.jpg
cat.jpg
ceiling.jpg
chair.jpg
cloth.jpg
computer.jpg
cow.jpg
cup.jpg
curtain.jpg
diningtable.jpg
dog.jpg
door.jpg
fence.jpg
floor.jpg
flower.jpg
food.jpg
grass.jpg
ground.jpg
horse.jpg
keyboard.jpg
light.jpg
motorbike.jpg
mountain.jpg
mouse.jpg
person.jpg
plate.jpg
platform.jpg
pottedplant.jpg
road.jpg
rock.jpg
sheep.jpg
shelves.jpg
sidewalk.jpg
sign.jpg
sky.jpg
snow.jpg
sofa.jpg
track.jpg
train.jpg
tree.jpg
truck.jpg
tvmonitor.jpg
wall.jpg
water.jpg
window.jpg
wood.jpg