ICLR 2022

LANGUAGE-DRIVEN SEMANTIC SEGMENTATION

Boyi Li
Cornell University, Cornell Tech

Kilian Q. Weinberger
Cornell University

Serge Belongie
University of Copenhagen

Vladlen Koltun
Apple

René Ranftl
Intel Labs

Mengxue

Motivation

- Text embeddings provide a flexible label representation in which semantically similar labels map to similar regions in the embedding space(e.g., “cat” and “furry”).

Text encoding

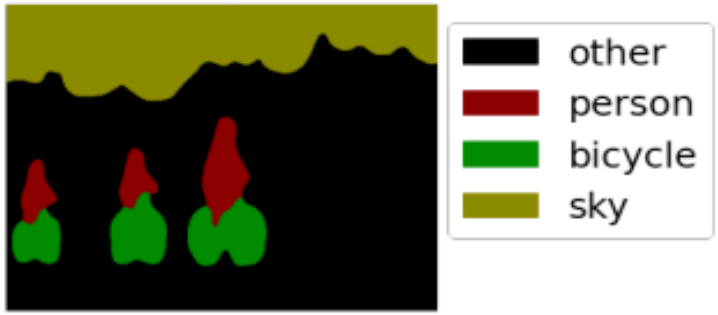
one-hot

person: 1 0 0 0…

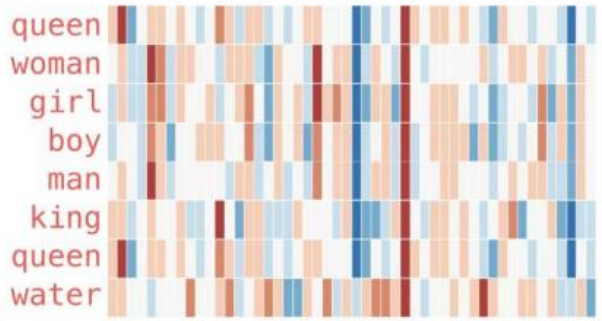
bicycle: 0 1 0 0 …

sky: 0 0 1 0 …

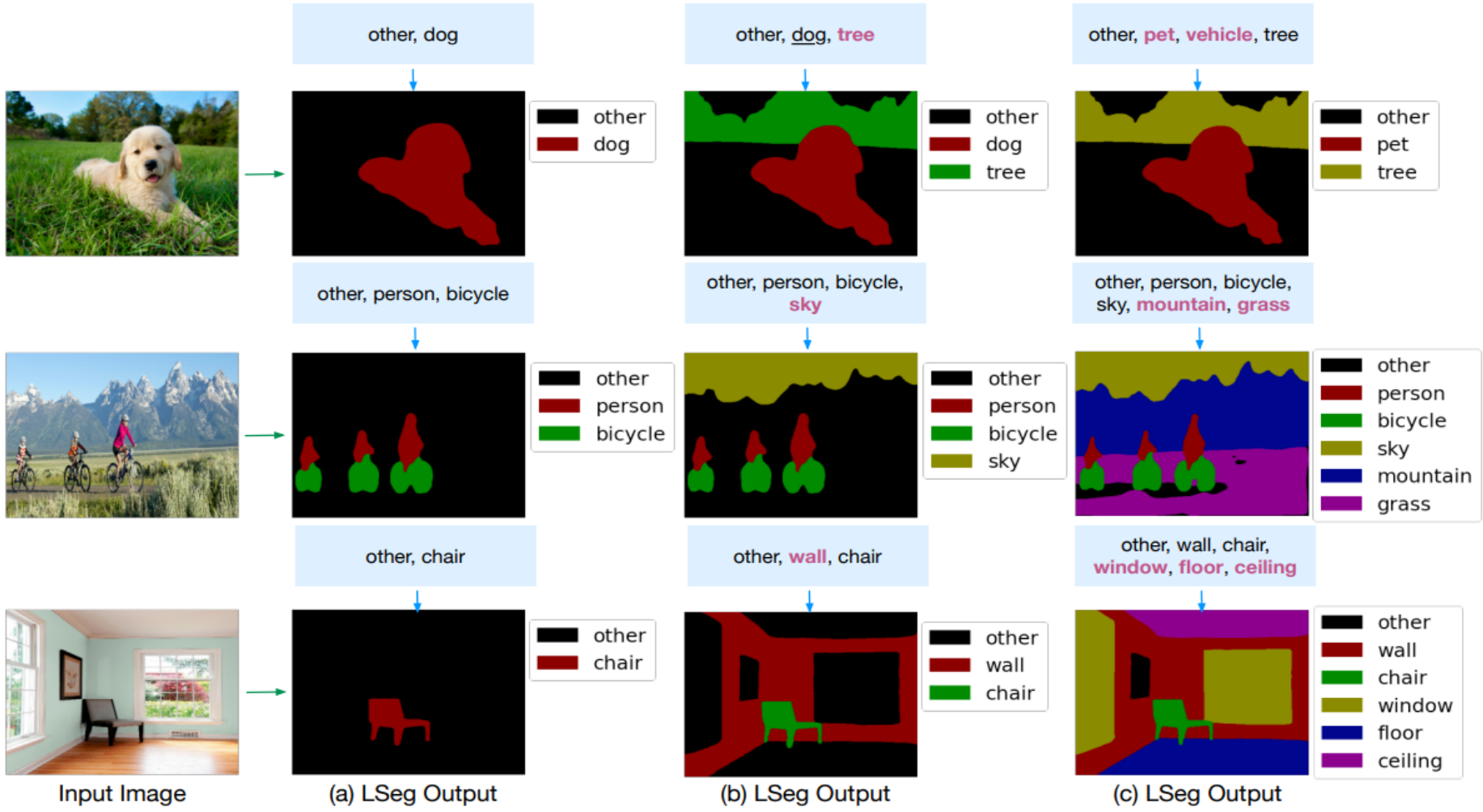
… …



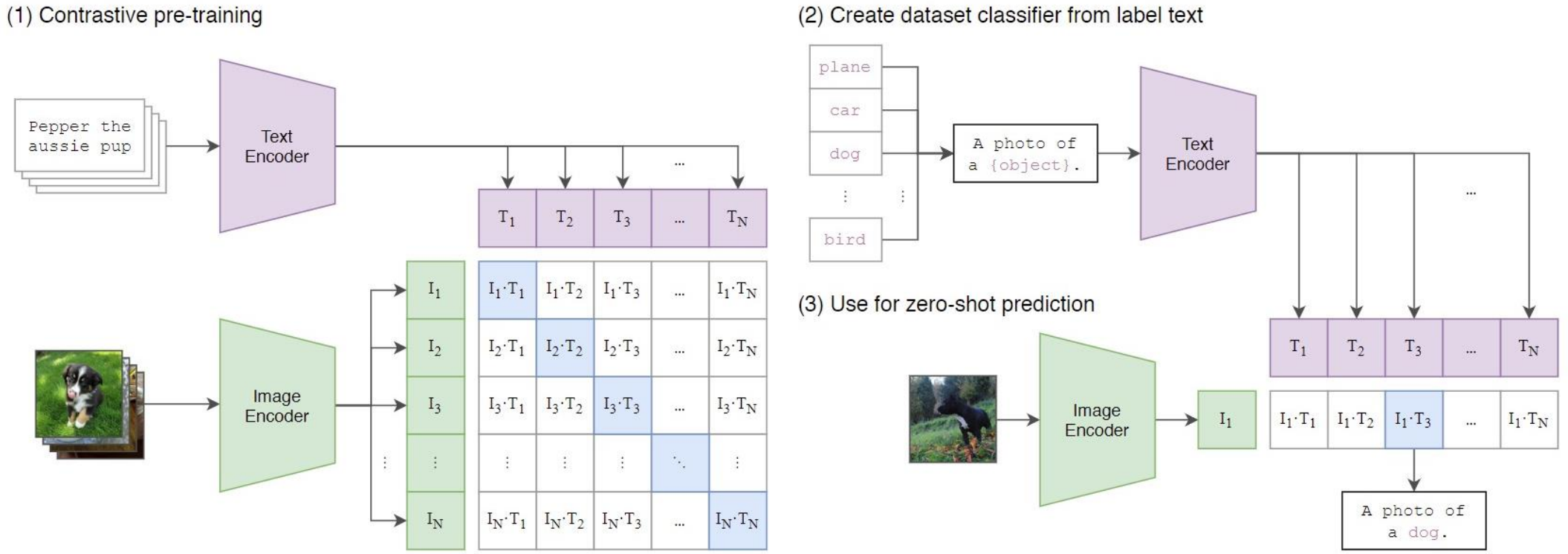
embedding



Contribution



Review--CLIP



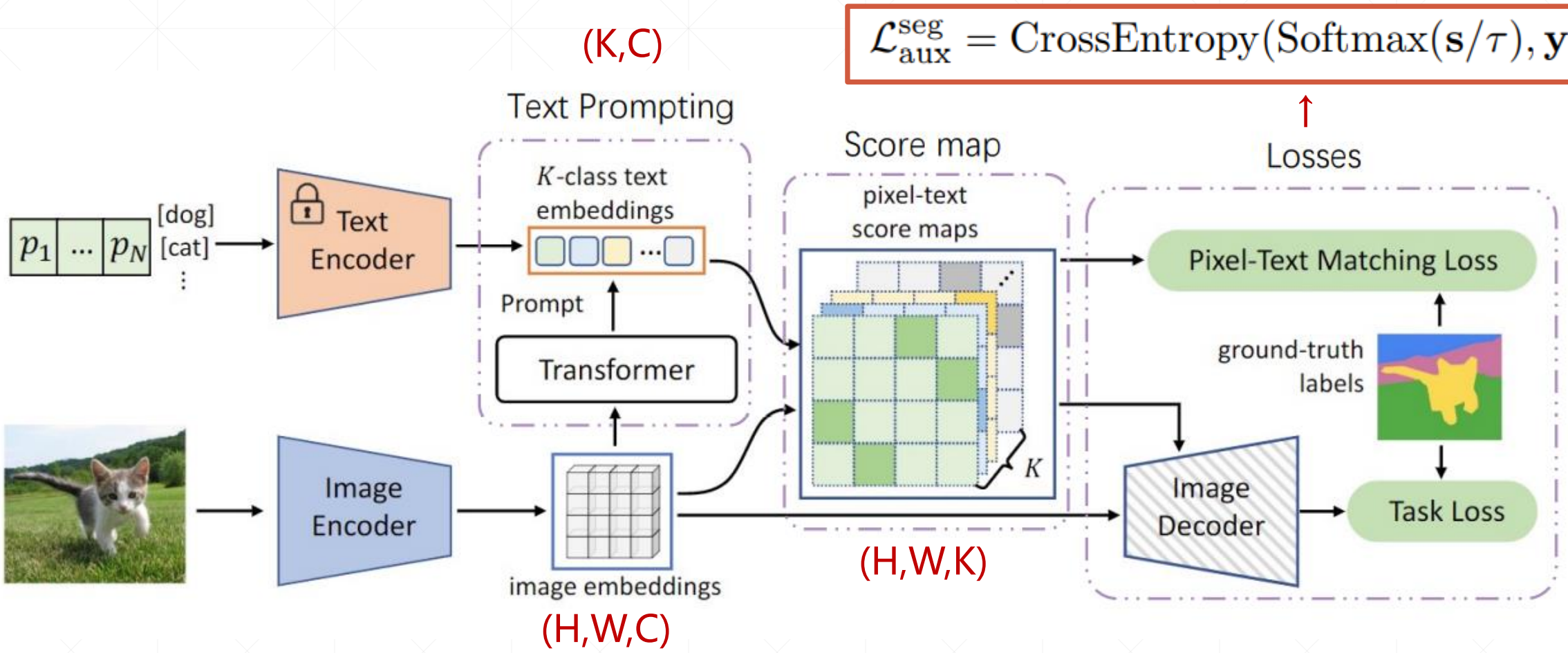
Positive pair

Negtive pair

→ InfoNCE

$$\mathop{\mathbb{E}}_{x,x^+,x^-}\left[-\log\left(\frac{e^{f(x)^T f(x^+)}}{e^{f(x)^T f(x^+)}+e^{f(x)^T f(x^-)}}\right)\right]$$

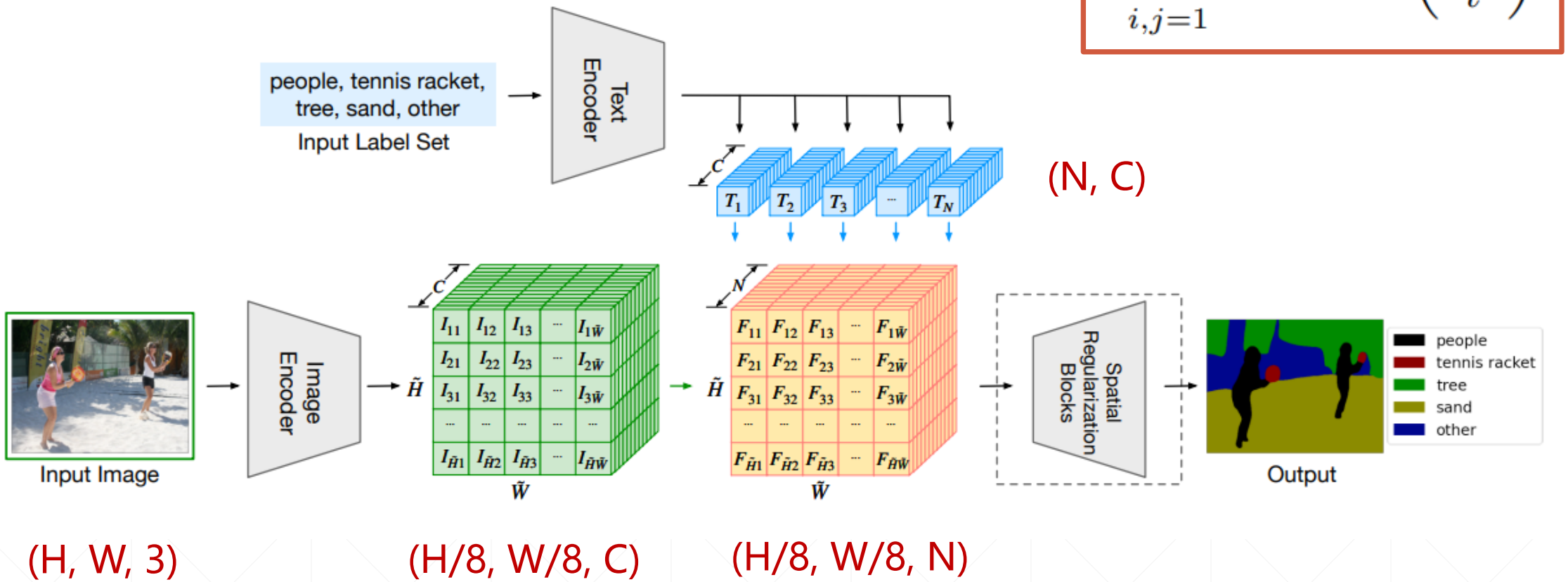
Review—DenseCLIP



CLIP Pre-training + Language-Guided Fine-tuning

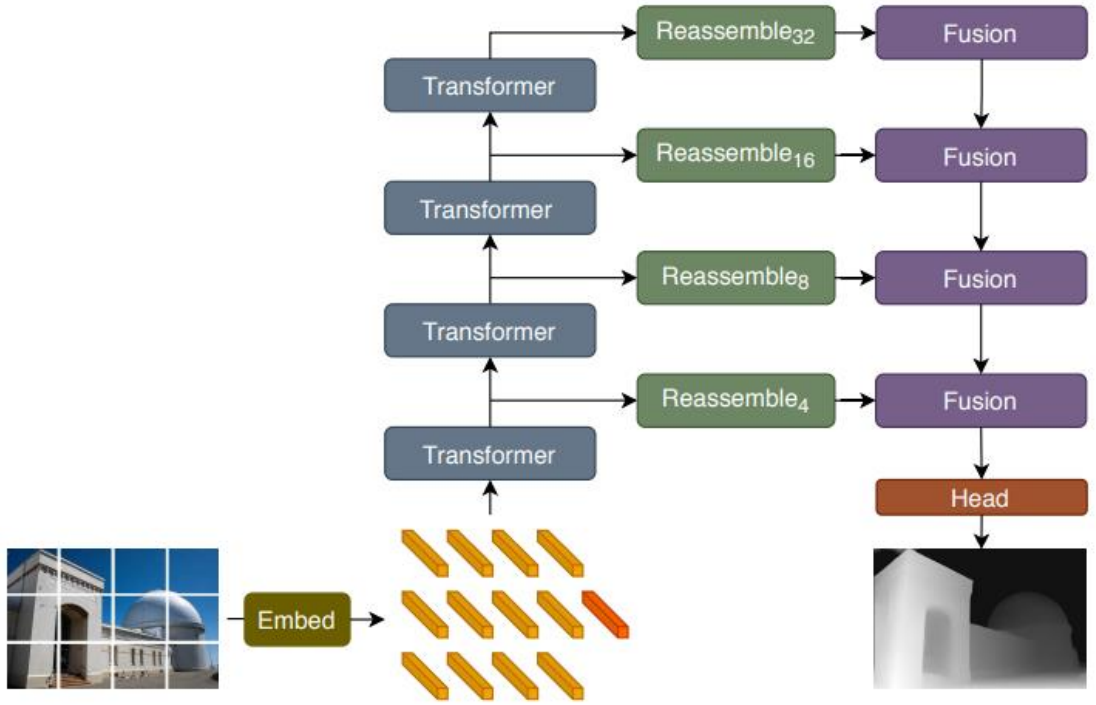
Approach — Framework

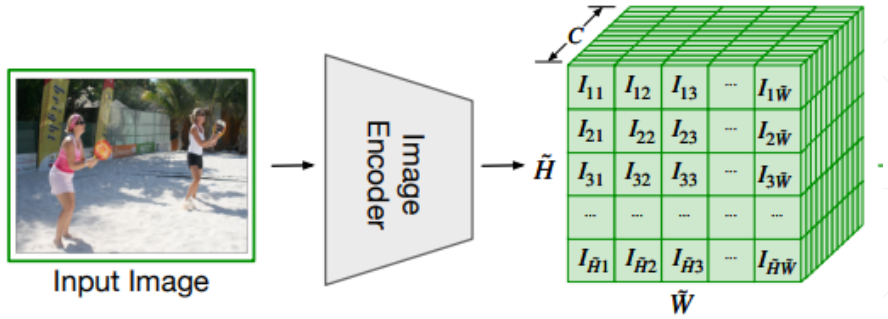
$$\sum_{i,j=1}^{H,W} \text{softmax}_{y_{ij}}\left(\frac{F_{ij}}{t}\right),$$



Approach — Image Encoder

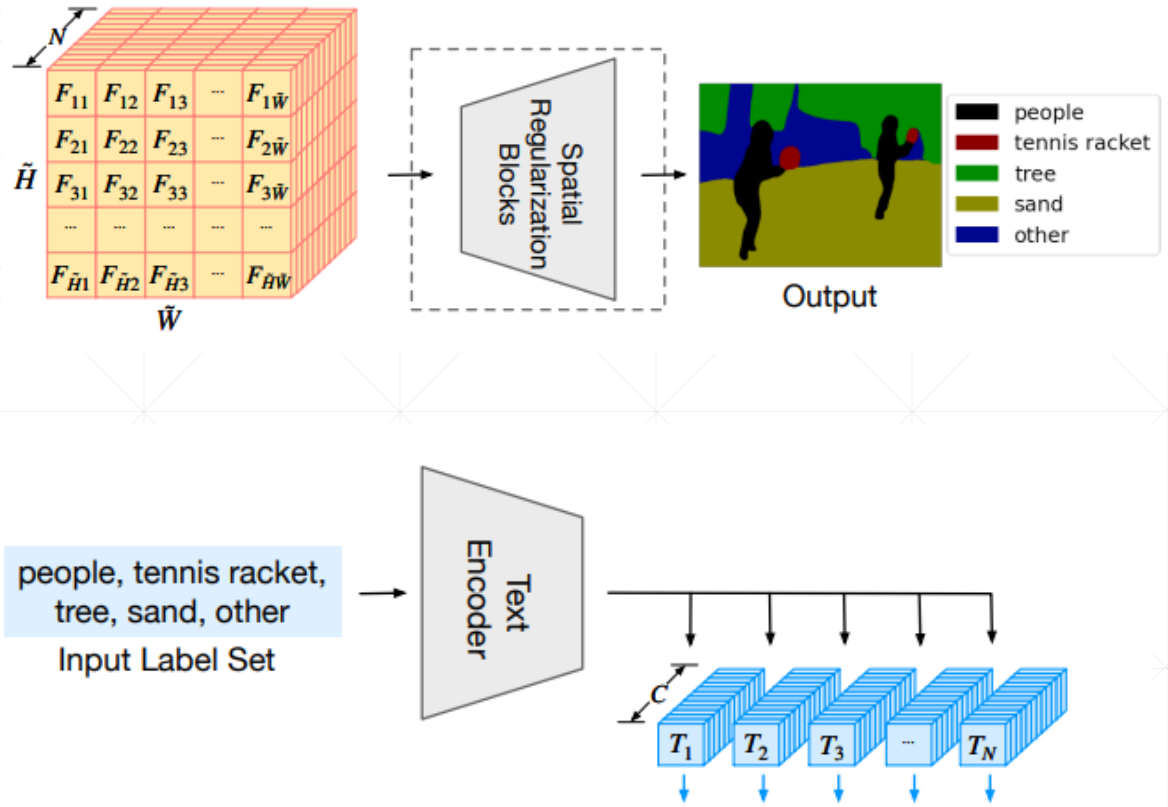
- Image Encoder = ViT image encoder + DPT decoder



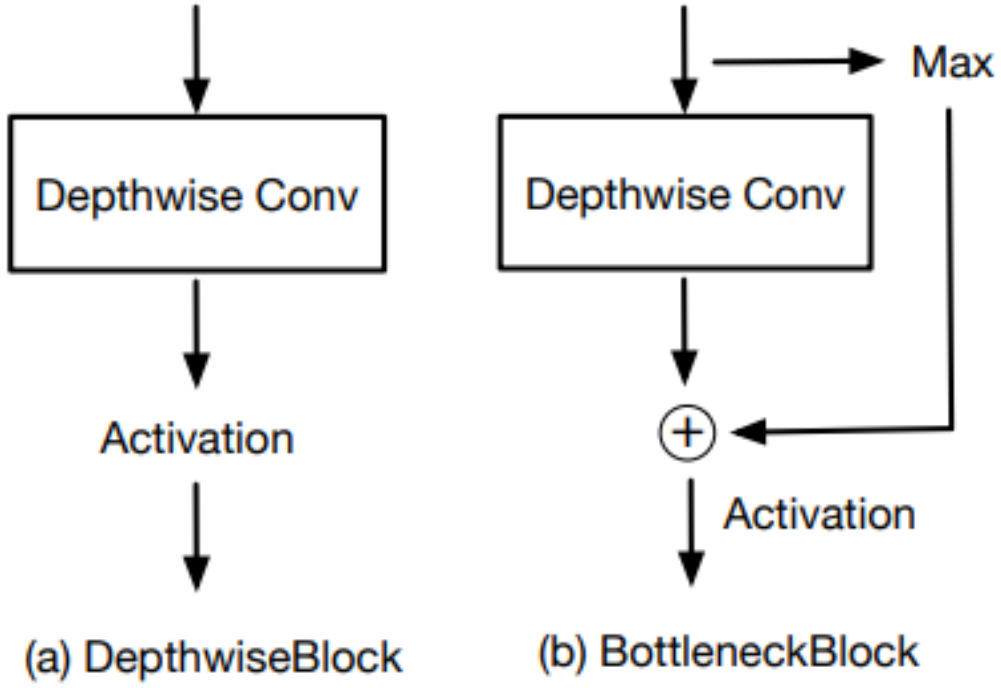


Approach—Spatial Regularization Blocks

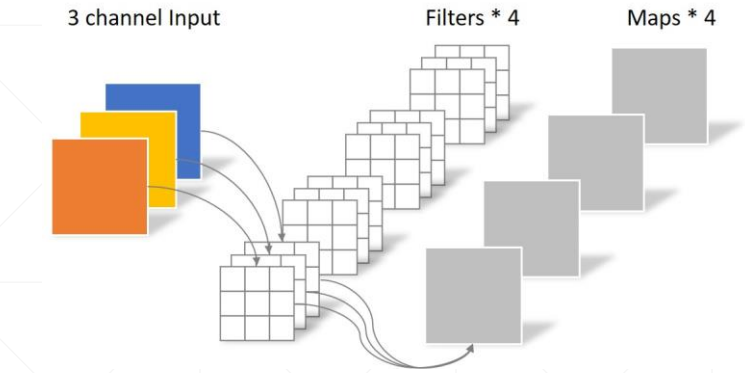
- Ensure all operations stay equivariant with respect to the labels
- There should be no interactions between the input channels, whose order is defined by the order of the words and can thus be arbitrary



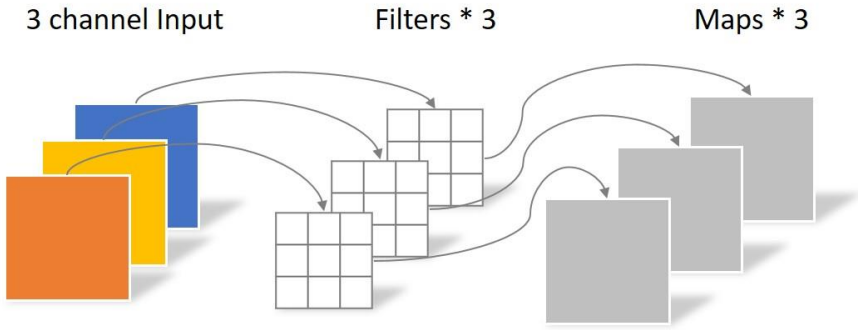
Approach—Spatial Regularization Blocks



Block Type	Metric	# depth 0	1	2	4
DepthwiseBlock	pixAcc [%]	79.70	79.72	**79.78**	7.67
	mIoU [%]	37.83	39.19	**39.45**	0.18
BottleneckBlock	pixAcc [%]	79.70	79.64	**79.70**	79.68
	mIoU [%]	37.83	39.16	**39.79**	38.78



Normal Convolution



Depthwise Convolution

Training details

- Image encoder: ViT/ResNet101 pretrained in ImageNet
  initialize the decoder of DPT randomly
- Text encoder: CLIP pretrained

- Freeze text encoder and only update the weights of the image encoder

- Provide the full label set that is defined by each training set to the text encoder for each image.

Experiments

- **Dataset** :
- PASCAL-5i
- COCO-20i
- FSSS-1000: train,val,test→520,240,240 classes
- **Evaluation Metric**
- mIoU: the average IoU over all classes
- FB-IoU: mean value of foreground and background IoUs in fold i and ignores the object classes

PASCAL-5i

label set index	Semantic Classes
0	aeroplane, bicycle, bird, boat, bottle
1	bus, car, cat, chair, cow
2	diningtable, dog, horse, motorbike, person
3	potted plant, sheep, sofa, train, tv/monitor

sub-dataset	train label set	val label set
0	1,2,3	0
1	0,2,3	1
2	0,1,3	2
3	0,1,2	3

COCO-20i

$COCO\text{-}20^0$	$COCO\text{-}20^1$	$COCO\text{-}20^2$	$COCO\text{-}20^3$
1 Person	2 Bicycle	3 Car	4 Motorcycle
5 Airplane	6 Bus	7 Train	8 Truck
9 Boat	10 T.light	11 Fire H.	12 Stop
13 Park meter	14 Bench	15 Bird	16 Cat
17 Dog	18 Horse	19 Sheep	20 Cow
21 Elephant	22 Bear	23 Zebra	24 Giraffe
25 Backpack	26 Umbrella	27 Handbag	28 Tie
29 Suitcase	30 Frisbee	31 Skis	32 Snowboard
33 Sports ball	34 Kite	35 B. bat	36 B. glove
37 Skateboard	38 Surfboard	39 T. racket	40 Bottle
41 W. glass	42 Cup	43 Fork	44 Knife
45 Spoon	46 Bowl	47 Banana	48 Apple
49 Sandwich	50 Orange	51 Broccoli	52 Carrot
53 Hot dog	54 Pizza	55 Donut	56 Cake
57 Chair	58 Couch	59 P. plant	60 Bed
61 D. table	62 Toilet	63 TV	64 Laptop
65 Mouse	66 Remote	67 Keyboard	68 Cellphone
69 Microwave	70 Oven	71 Toaster	72 Sink
73 Fridge	74 Book	75 Clock	76 Vase
77 Scissors	78 Teddy	79 Hairdrier	80 Toothbrush

Experiments

Model	Backbone	Method	5^0	5^1	5^2	5^3	mean	FB-IoU
OSLSM	VGG16	1-shot	33.6	55.2	40.9	33.5	40.8	61.3
co-FCN		1-shot	36.7	50.6	44.9	32.4	41.1	60.1
AMP-2		1-shot	41.9	50.2	46.7	34.7	43.4	61.9
PANet	ResNet50	1-shot	44.0	57.5	50.8	44.0	49.1	-
PGNet		1-shot	56.0	66.9	50.6	50.4	56.0	69.9
FWB	ResNet101	1-shot	51.3	64.5	56.7	52.2	56.2	-
PPNet		1-shot	52.7	62.8	57.4	47.7	55.2	70.9
DAN		1-shot	54.7	68.6	57.8	51.6	58.2	71.9
PFENet		1-shot	60.5	69.4	54.4	55.9	60.1	72.9
RePRI		1-shot	59.6	68.6	**62.2**	47.2	59.4	-
HSNet		1-shot	**67.3**	**72.3**	62.0	**63.1**	**66.2**	**77.6**
SPNet	ResNet101	zero-shot	23.8	17.0	14.1	18.3	18.3	44.3
ZS3Net		zero-shot	40.8	39.4	39.3	33.6	38.3	57.7
LSeg	ResNet101	zero-shot	**52.8**	**53.8**	**44.4**	**38.5**	**47.4**	**64.1**
LSeg	ViT-L/16	zero-shot	**61.3**	**63.6**	**43.1**	**41.0**	**52.3**	**67.0**

Table 1: Comparison of mIoU and FB-IoU (higher is better) on PASCAL-5^i.

Experiments

Model	Backbone	Method	20^0	20^1	20^2	20^3	mean	FB-IoU
PPNet	ResNet50	1-shot	28.1	30.8	29.5	27.7	29.0	-
PMM		1-shot	29.3	34.8	27.1	27.3	29.6	-
RPMM		1-shot	29.5	36.8	28.9	27.0	30.6	-
RePRI		1-shot	32.0	38.7	32.7	33.1	34.1	-
FWB	ResNet101	1-shot	17.0	18.0	21.0	28.9	21.2	-
DAN		1-shot	-	-	-	-	24.4	62.3
PFENet		1-shot	36.8	41.8	38.7	36.7	38.5	63.0
HSNet		1-shot	**37.2**	**44.1**	**42.4**	**41.3**	**41.2**	**69.1**
ZS3Net	ResNet101	zero-shot	18.8	20.1	24.8	20.5	21.1	55.1
LSeg	ResNet101	zero-shot	**22.1**	**25.1**	**24.9**	**21.5**	**23.4**	**57.9**
LSeg	ViT-L/16	zero-shot	**28.1**	**27.5**	**30.0**	**23.2**	**27.2**	**59.9**

Table 2: Comparison of mIoU and FB-IoU (higher is better) on COCO-20^i.

Experiments

Table 3 compares our approach to state-of-the-art few-shot models. Notably, under the same ResNet101, LSeg could achieve comparative results of the state-of-the-art one-shot method. Also, LSeg even outperforms a state-of-the-art one-shot method: 87.8 mIoU (ours) vs. 86.5 mIoU (HSNet) with a larger backbone ViT-L/16, indicating that LSeg generalizes very well to unseen categories. Figure 4 shows examples of segmentation results on unseen categories.

Model	Backbone	Method	mIoU
OSLSM	VGG16	1-shot	70.3
GNet		1-shot	71.9
FSS		1-shot	73.5
DoG-LSTM		1-shot	80.8
DAN	ResNet101	1-shot	85.2
HSNet		1-shot	86.5
LSeg	ResNet101	zero-shot	84.7
LSeg	ViT-L/16	zero-shot	**87.8**

Table 3: Comparison of mIoU on FSS-1000.

Visualization

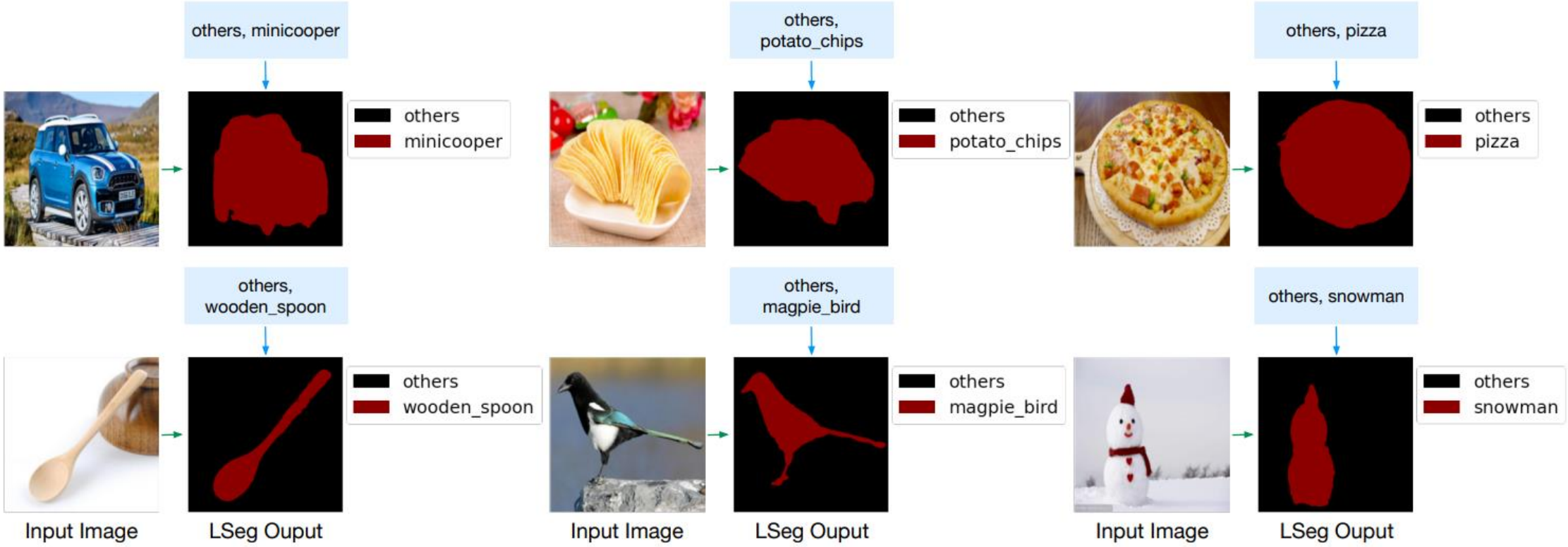


Figure 4: LSeg zero-shot semantic segmentation results on unseen categories of FSS-1000 dataset.

Experiments

Method	Backbone	Text Encoder	pixAcc [%]	mIoU [%]
OCNet	ResNet101	-	-	45.45
ACNet	ResNet101	-	81.96	45.90
DeeplabV3	ResNeSt101	-	82.07	46.91
DPT	ViT-L/16	-	**82.70**	**47.63**
LSeg	ViT-L/16	ViT-B/32	82.46	46.28
LSeg	ViT-L/16	RN50 $\times$ 16	**82.78**	**47.25**

Table 6: Comparison of semantic segmentation results on the ADE20K validation set. For LSeg, we conduct experiments with fixed text encoders of ViT-B/32 and RN50 $\times$ 16 CLIP pretrained models.

Visualization

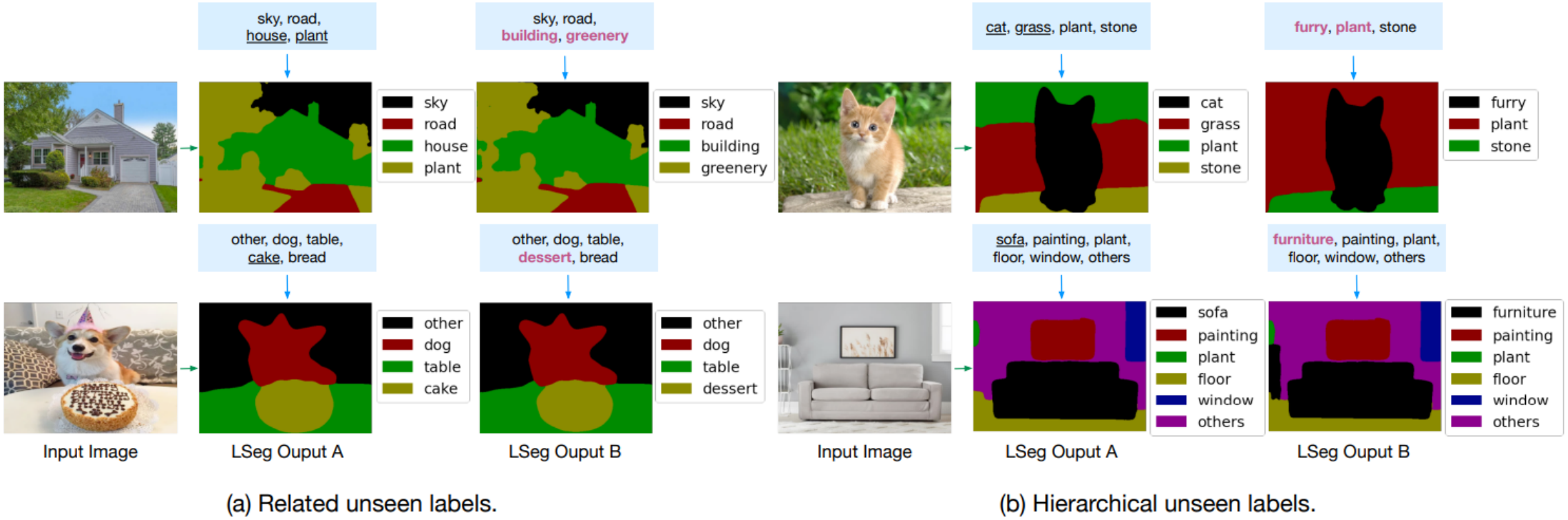


Figure 5: LSeg examples with related but previously unseen labels, and hierarchical labels. Going from left to right, labels that are removed between runs are underlined, whereas labels that are added are marked in **bold red**.

Thanks