Open-Vocabulary Segmentation

Presented by

Saikat Dutta

PhD Scholar

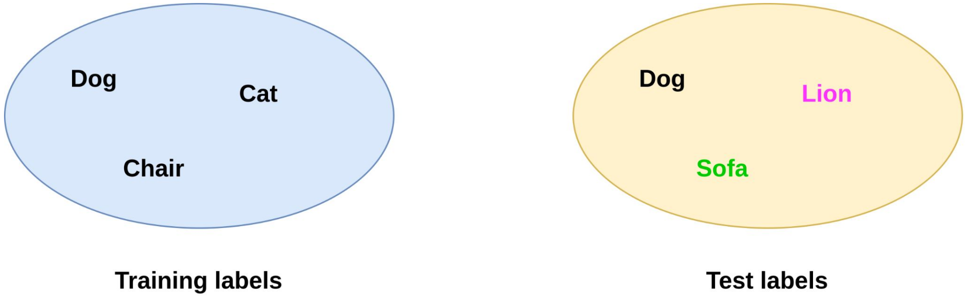
IITB-Monash Research Academy

Outline

- ➢ Introduction
- ➢ Basics: CLIP
- ➢ Language-Driven Semantic Segmentation (LSeg)
- ➢ CAT-Seg: Cost Aggregation for Open-Vocabulary Semantic Segmentation
- ➢ Q/A

Introduction

➢ Open-Vocabulary Segmentation (OVS) aims to segment objects from an open-set of categories.
  - Training labels ≠ Test labels
  - Zero-shot transfer!

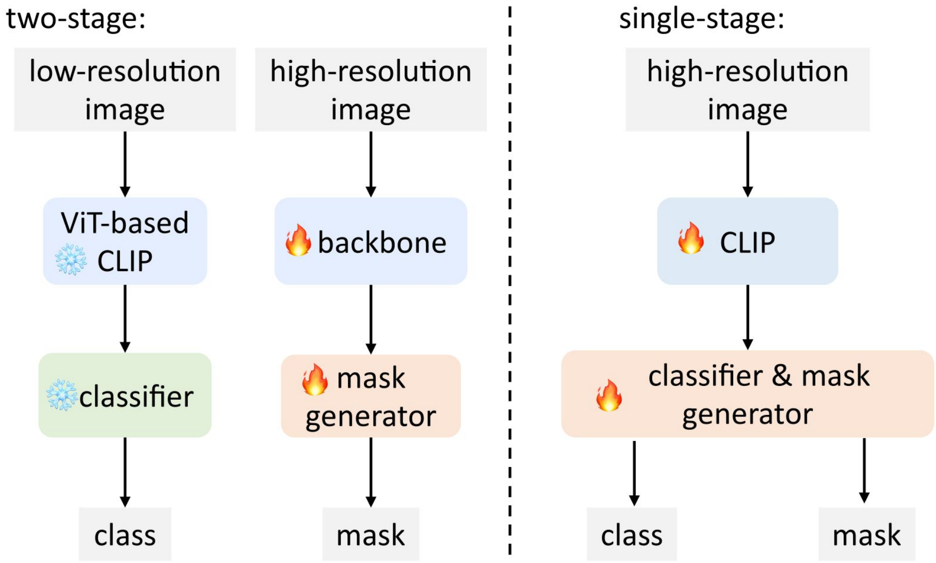


➢ Alignment between visual and semantic feature space in visual-language models (VLMs) e.g. CLIP is often exploited for OVS.

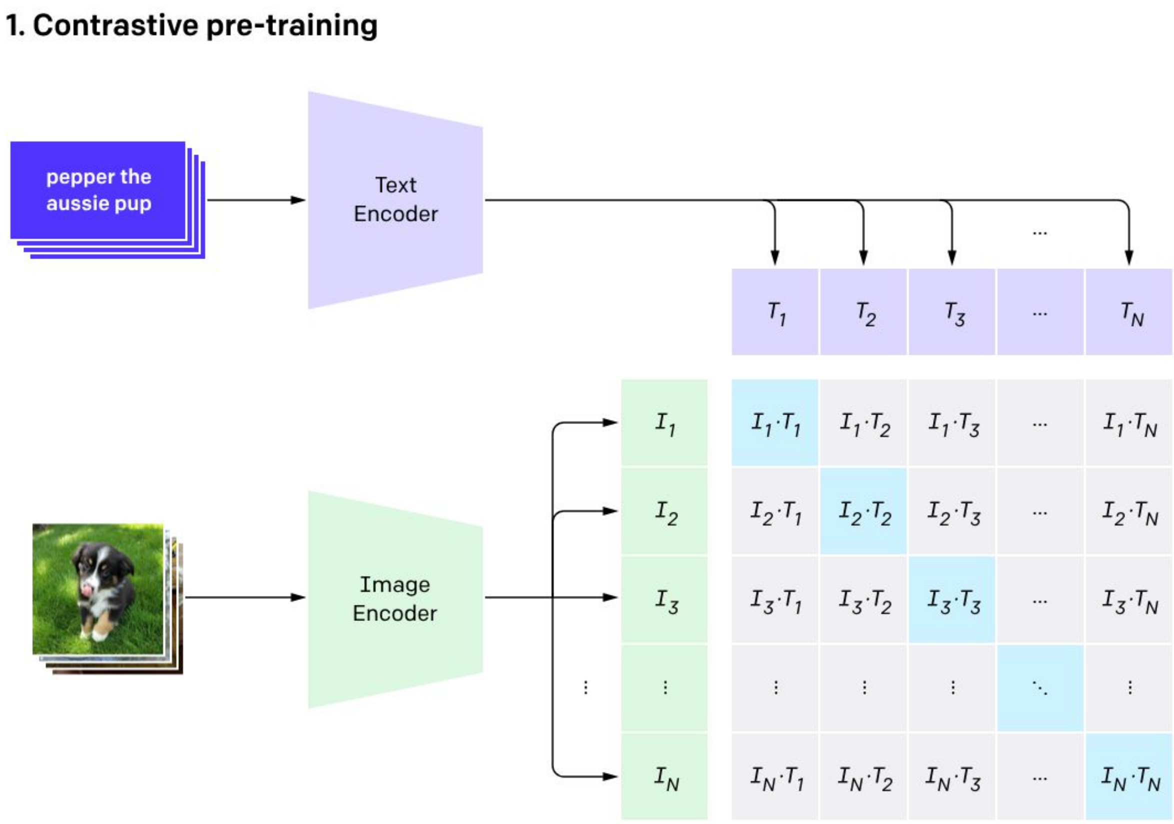
Introduction

➢ **Types of OVS methods:**

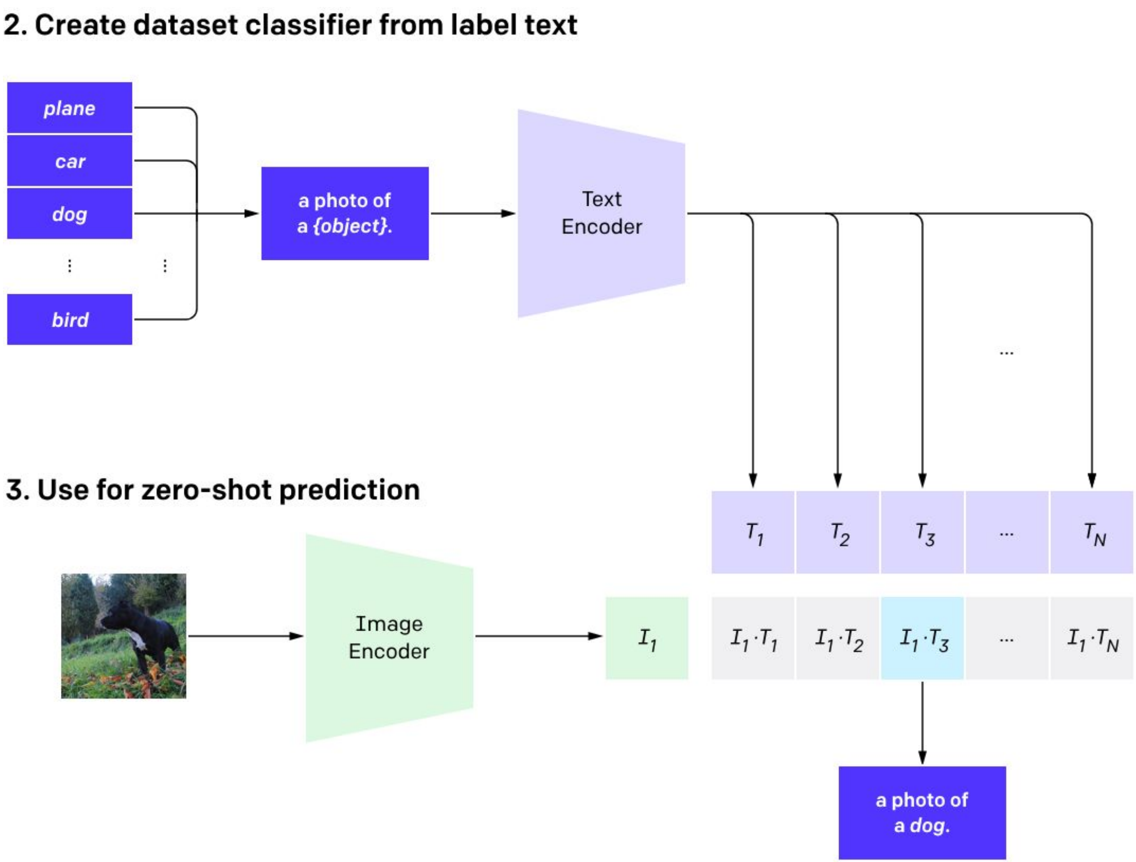
- **Two-stage approaches:**
  - First predict class-agnostic region proposals then feed them to CLIP for final predictions.
- **One-stage approaches:**
  - Embeddings from CLIP are directly used to predict masks.



Basics: CLIP


1. Contrastive pre-training
pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Basics: CLIP

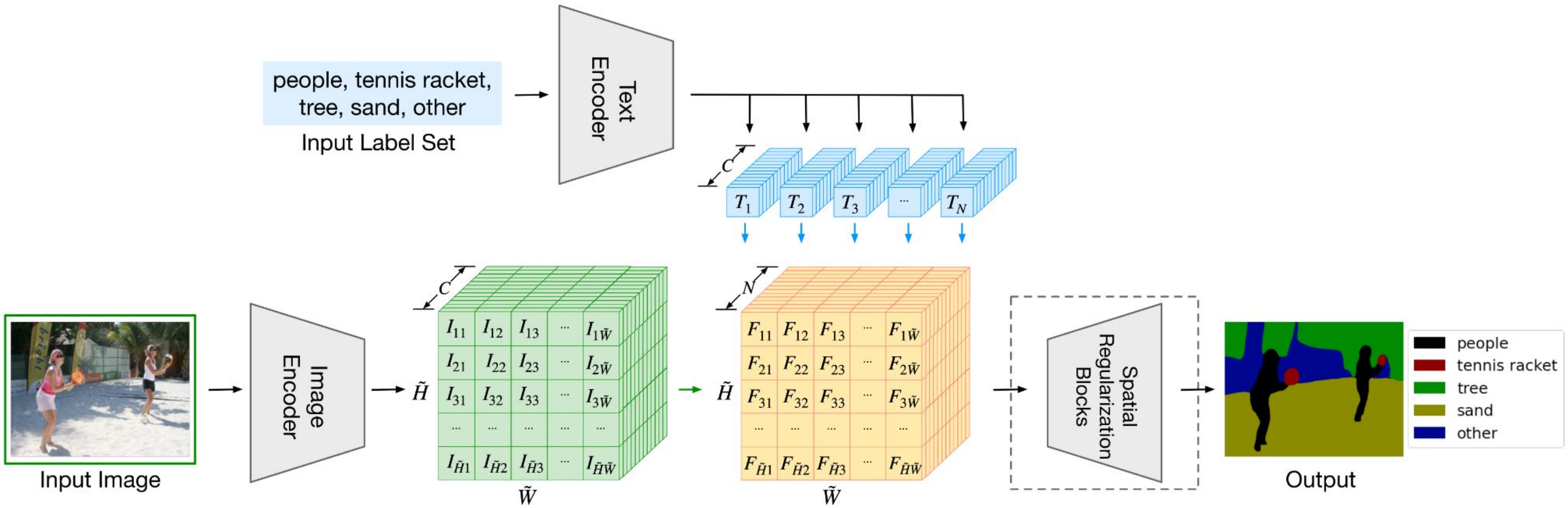


2. Create dataset classifier from label text
plane
car
dog
bird
a photo of a {object}.
Text Encoder
3. Use for zero-shot prediction
Image Encoder
I_1
T_1
T_2
T_3
...
T_N
$I_1 \cdot T_1$
$I_1 \cdot T_2$
$I_1 \cdot T_3$
$I_1 \cdot T_N$
a photo of a dog.

Language-Driven Semantic Segmentation (LSeg)

ICLR 2022

Language-Driven Semantic Segmentation (LSeg)



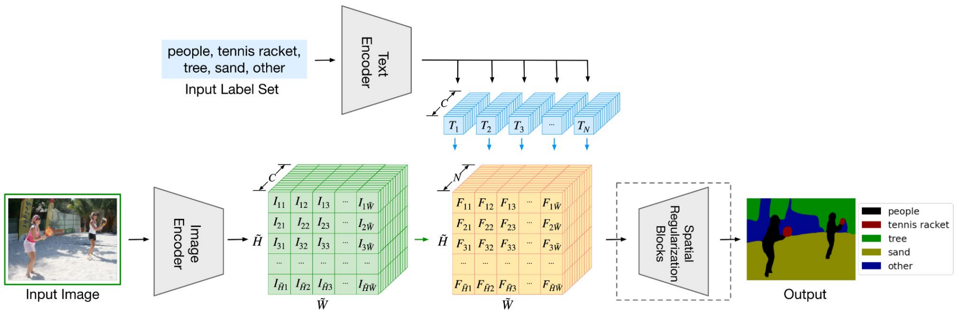
Language-Driven Semantic Segmentation (LSeg)

- **Text Encoder:** CLIP Text encoder
- **Image Encoder:** ViT
- **Word-pixel correlation tensor:**

$$f_{ijk} = I_{ij} \cdot T_k.$$

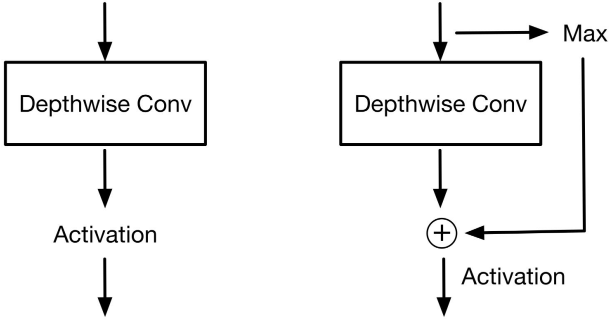
$$F_{ij} \in \mathbb{R}^N$$

$$F_{ij} = (f_{ij1}, f_{ij2}, ..., f_{ijk})^T$$



Language-Driven Semantic Segmentation (LSeg)

- **Spatial Regularization Block:**
  - Decoding block
  - Ensure: all operations need to stay equivariant w.r.t. labels
  - Depthwise convolution is used.



(a) DepthwiseBlock (b) BottleneckBlock

Language-Driven Semantic Segmentation (LSeg)

- **Quantitative Results:**

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

Language-Driven Semantic Segmentation (LSeg)

- **Qualitative Results:**

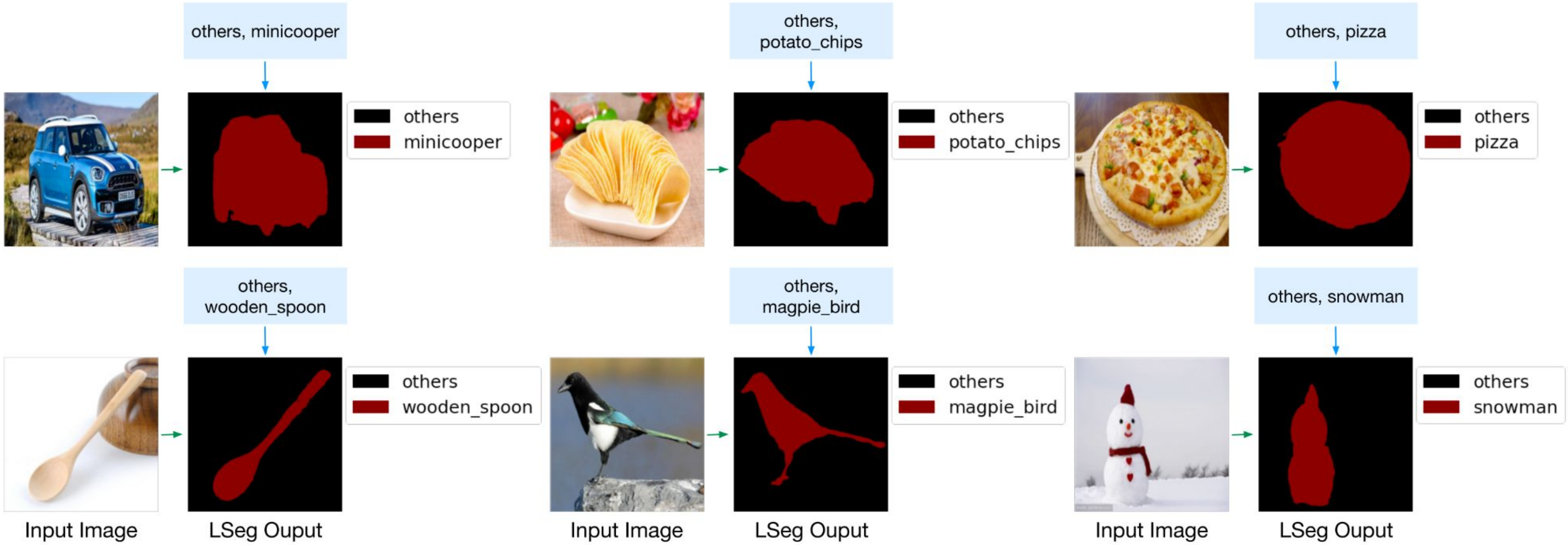


Figure 4: LSeg zero-shot semantic segmentation results on unseen categories of FSS-1000 dataset.

Language-Driven Semantic Segmentation (LSeg)

- **Qualitative Results:**

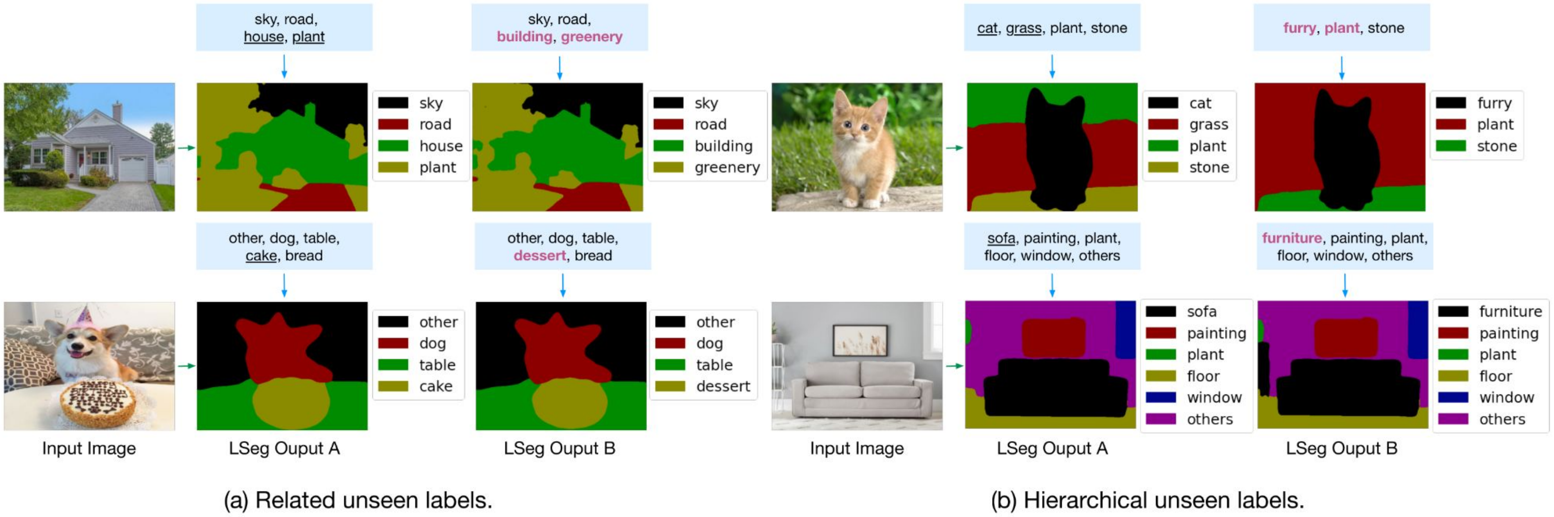
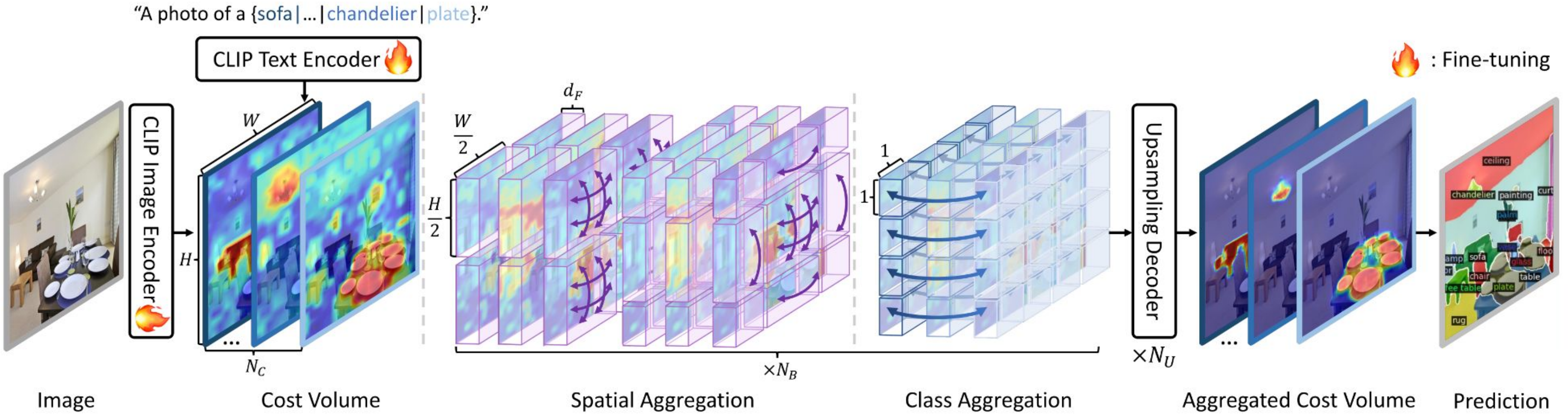


Figure 5: LSeg examples with related but previously unseen labels, and hierarchical labels. Going from left to right, labels that are removed between runs are <u>underlined</u>, whereas labels that are added are marked in **bold red**.

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

CVPR 2024

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



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

- **Cost Computation and Embedding:**

Image Embeddings: $D^V = \Phi^V(I) \in \mathbb{R}^{(H\times W)\times d}$

Text Embeddings: $D^L = \Phi^L(T) \in \mathbb{R}^{N_\mathcal{C}\times d}$

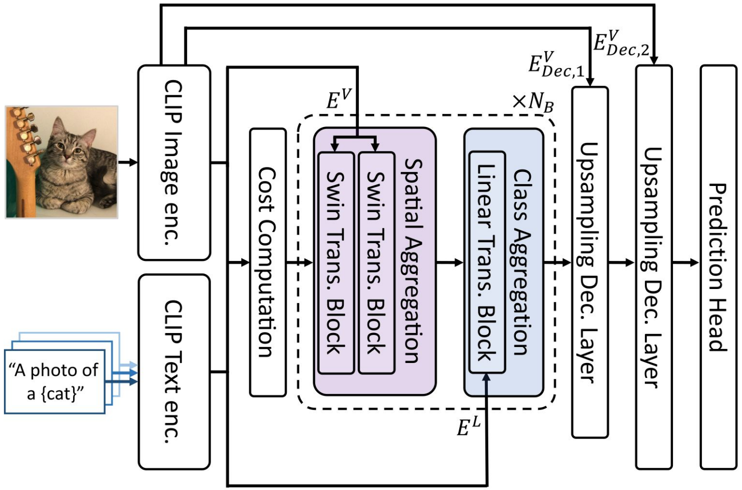
Cost: $C(i,n) = \frac{D^V(i) \cdot D^L(n)}{\|D^V(i)\| \|D^L(n)\|}$. $C \in \mathbb{R}^{(H\times W)\times N_\mathcal{C}}$

Cost Volume: $F \in \mathbb{R}^{(H\times W)\times N_\mathcal{C}\times d_F}$

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

- **Spatial Aggregation:**
  - Operates on each class slice separately
  - Consists of Swin-T blocks
  - Embedding guidance is used from CLIP vision embeddings.

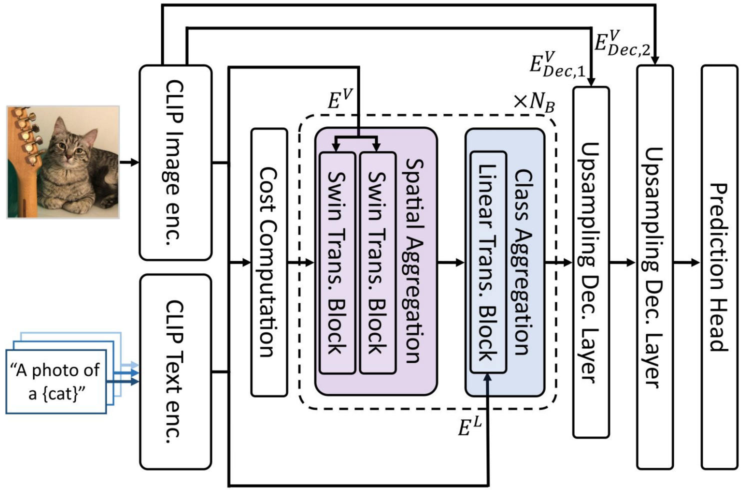
$$F'(:,n) = \mathcal{T}^{\mathrm{sa}}([F(:,n); \mathcal{P}^V(D^V)])$$



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

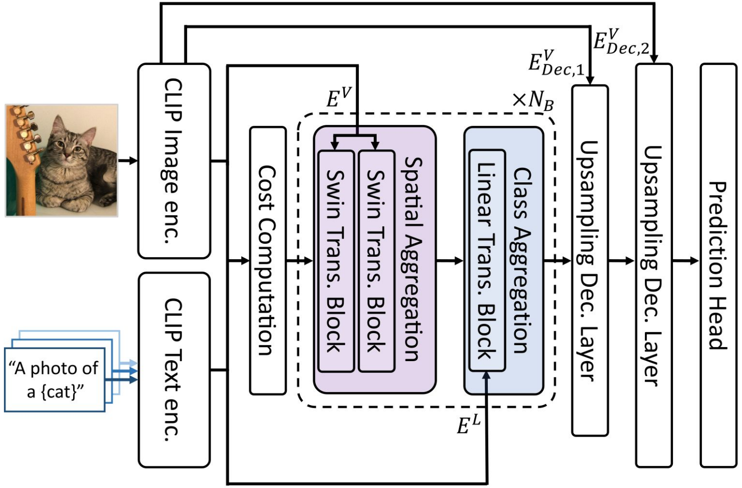
- **Class Aggregation:**
  - Operates on each spatial location separately
  - Consists of Linear Transformer blocks
  - Embedding guidance is used from CLIP text embeddings.

$$F''(i,:) = \mathcal{T}^{\text{ca}}([F'(i,:); \mathcal{P}^{L}(D^{L})]),$$



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

- **Upsampling Decoder:**
  - Standard decoding block consisting of:
    - Bilinear Upsampling
    - Concatenation with intermediate features from CLIP Vision encoder
    - Conv layer



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

- **Quantitative Results:**

Model	VLM	Additional Backbone	Training Dataset	Additional Dataset	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b
SPNet [82]	-	ResNet-101	PASCAL VOC	✗	-	-	-	24.3	18.3	-
ZS3Net [6]	-	ResNet-101	PASCAL VOC	✗	-	-	-	19.4	38.3	-
LSeg [40]	CLIP ViT-B/32	ResNet-101	PASCAL VOC-15	✗	-	-	-	-	47.4	-
LSeg+ [22]	ALIGN	ResNet-101	COCO-Stuff	✗	2.5	5.2	13.0	36.0	-	59.0
ZegFormer [15]	CLIP ViT-B/16	ResNet-101	COCO-Stuff-156	✗	4.9	9.1	16.9	42.8	86.2	62.7
ZegFormer† [15]	CLIP ViT-B/16	ResNet-101	COCO-Stuff	✗	5.6	10.4	18.0	45.5	89.5	<u>65.5</u>
ZSseg [84]	CLIP ViT-B/16	ResNet-101	COCO-Stuff	✗	7.0	-	20.5	47.7	88.4	-
OpenSeg [22]	ALIGN	ResNet-101	COCO Panoptic	✓	4.4	7.9	17.5	40.1	-	63.8
OVSeg [43]	CLIP ViT-B/16	ResNet-101c	COCO-Stuff	✓	7.1	11.0	24.8	53.3	92.6	-
ZegCLIP [100]	CLIP ViT-B/16	-	COCO-Stuff-156	✗	-	-	-	41.2	93.6	-
SAN [85]	CLIP ViT-B/16	-	COCO-Stuff	✗	<u>10.1</u>	<u>12.6</u>	<u>27.5</u>	<u>53.8</u>	<u>94.0</u>	-
CAT-Seg (ours)	CLIP ViT-B/16	-	COCO-Stuff	✗	**12.0** (+1.9)	**19.0** (+6.4)	**31.8** (+4.3)	**57.5** (+3.7)	**94.6** (+0.6)	**77.3** (+11.8)
LSeg [40]	CLIP ViT-B/32	ViT-L/16	PASCAL VOC-15	✗	-	-	-	-	52.3	-
OpenSeg [22]	ALIGN	Eff-B7	COCO Panoptic	✓	8.1	11.5	26.4	44.8	-	<u>70.2</u>
OVSeg [43]	CLIP ViT-L/14	Swin-B	COCO-Stuff	✓	9.0	12.4	29.6	55.7	94.5	-
SAN [85]	CLIP ViT-L/14	-	COCO-Stuff	✗	<u>12.4</u>	<u>15.7</u>	<u>32.1</u>	<u>57.7</u>	<u>94.6</u>	-
ODISE [83]	CLIP ViT-L/14	Stable Diffusion	COCO-Stuff	✗	11.1	14.5	29.9	57.3	-	-
CAT-Seg (ours)	CLIP ViT-L/14	-	COCO-Stuff	✗	**16.0** (+3.6)	**23.8** (+8.1)	**37.9** (+5.8)	**63.3** (+5.6)	**97.0** (+2.4)	**82.5** (+12.3)

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

- **Qualitative Results:**

(a) SAN (b) **Ours** (c) GT

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

- **Ablation studies:**

	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**

Table 4. **Ablation study for CAT-Seg.** We conduct ablation study by gradually adding components to the cost aggregation baseline.

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

- **Ablation studies:**

	Methods	A-847	PC-459	A-150	PC-59	PAS-20	PAS-20^b	#param. (M)	Memory (GiB)
(I)	Freeze	10.4	15.0	31.8	52.5	92.2	71.3	5.8	20.0
(II)	Prompt	8.8	14.3	30.5	55.8	93.2	74.7	7.0	20.9
(III)	Full F.T.	13.6	22.2	34.0	61.1	**97.3**	79.7	393.2	26.8
(IV)	Attn. F.T.	15.7	23.7	37.1	63.1	97.1	81.5	134.9	20.9
(V)	QK F.T.	15.3	23.0	36.3	62.0	95.9	81.9	70.3	20.9
(VI)	KV F.T.	**16.1**	**23.8**	37.6	62.4	96.7	82.0	70.3	20.9
(VII)	QV F.T. (Img.)	13.9	22.8	35.1	62.0	96.3	82.0	56.7	20.9
(VIII)	QV F.T. (Txt.)	14.7	22.2	35.1	60.0	95.8	80.3	19.9	20.0
(IX)	QV F.T. (Both)	16.0	**23.8**	**37.9**	**63.3**	97.0	**82.5**	70.3	20.9

Table 6. **Analysis of fine-tuning methods for CLIP.** We additionally note the number of learnable parameters of CLIP and memory consumption during training. Our method not only outperforms full fine-tuning, but also requires smaller computation.

Thank you!

Questions?