Learning Hierarchical Image Segmentation For Recognition and By Recognition

Tsung-Wei Ke*

Sangwoo Mo*

Stella Yu

Berkeley UNIVERSITY OF CALIFORNIA

UNIVERSITY OF MICHIGAN

*Equal Contribution

ICLR

Image understanding happens at multiple levels: Image recognition vs. Pixel segmentation

Image | Part-level segmentation | Object-level segmentation | Image-level recognition

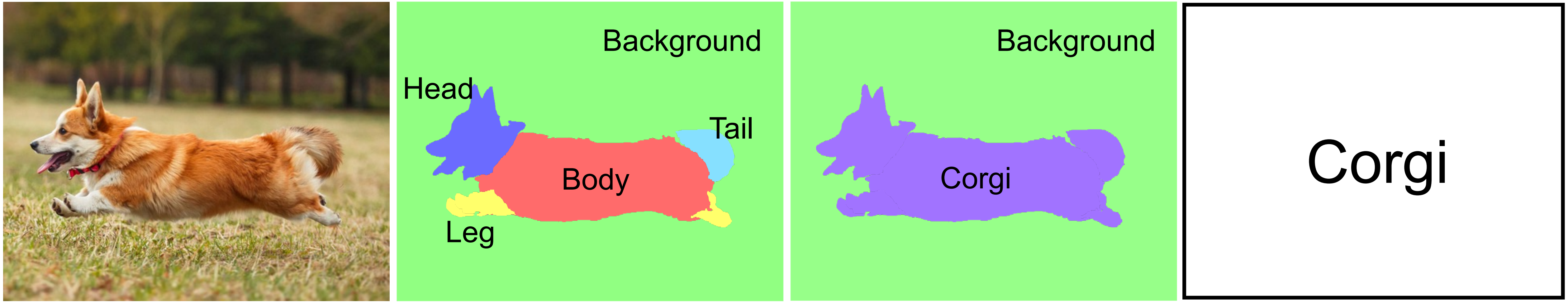
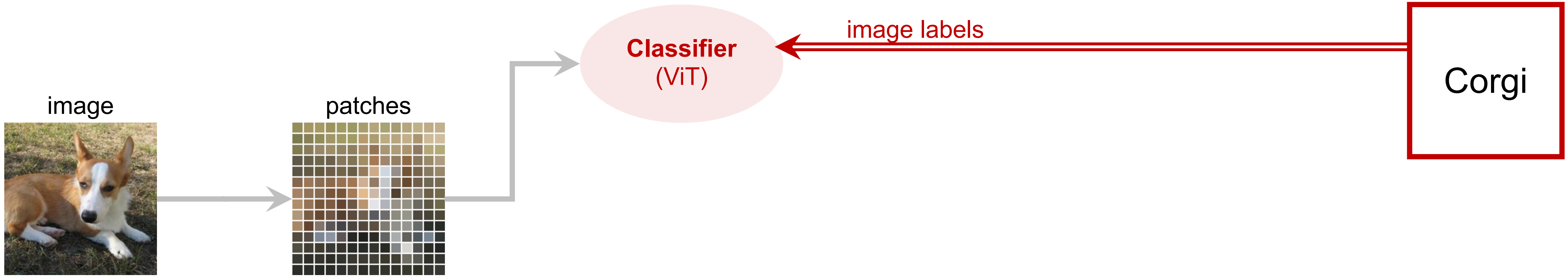


Image Credit: https://segment-anything.com/demo#

Prior works: Build separate models and supervision

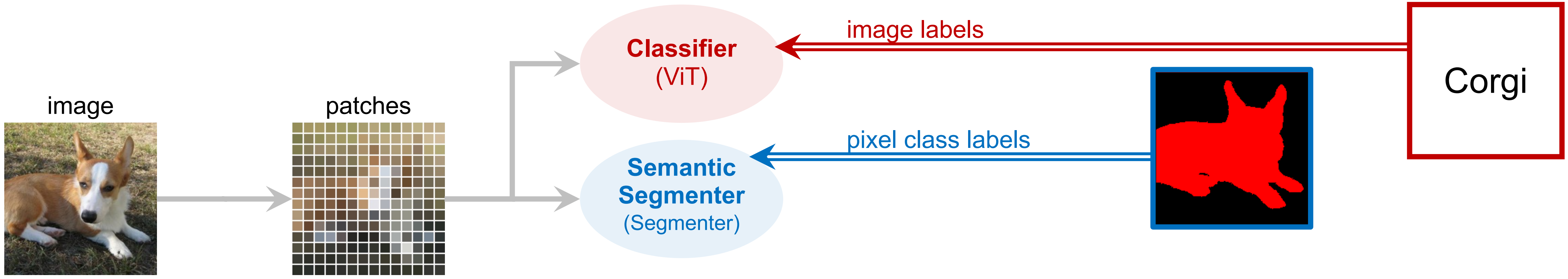


Imagenet: A large-scale hierarchical image database. Deng et al. *CVPR 2009*

Sun database: Large-scale scene recognition from abbey to zoo. Xiao et al. CVPR 2010

Laion-5b: An open large-scale dataset for training next generation image-text models. Schuhmann et al. *NeuRIPS 2022*

Prior works: Build separate models and supervision

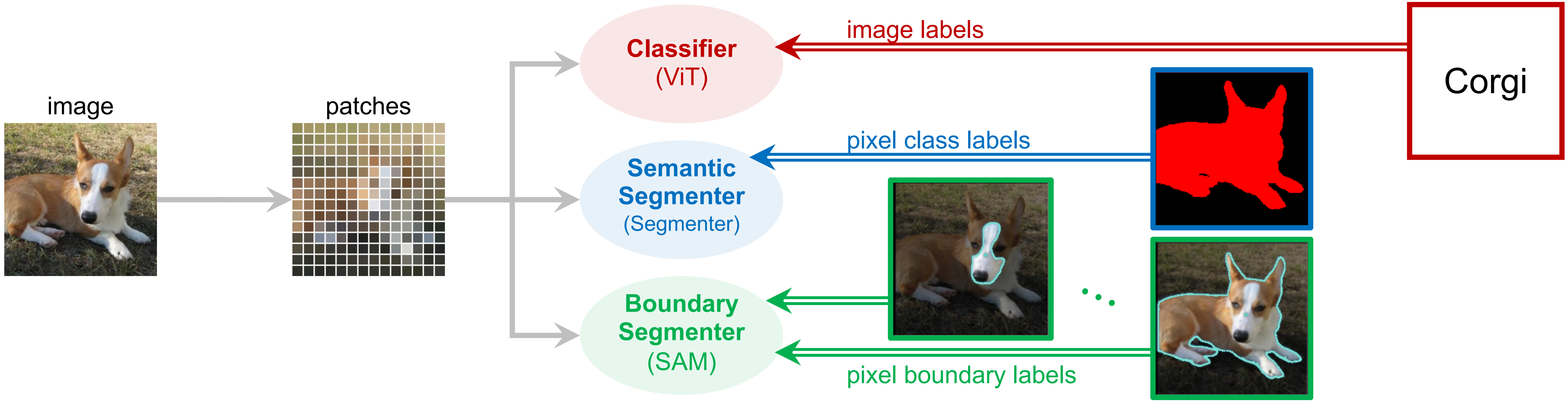


Microsoft coco: Common objects in context. Lin, et al. ECCV 2014

The cityscapes dataset for semantic urban scene understanding. Cordts et al. CVPR 2016

Scene parsing through ade20k dataset. Zhou et al. CVPR 2017

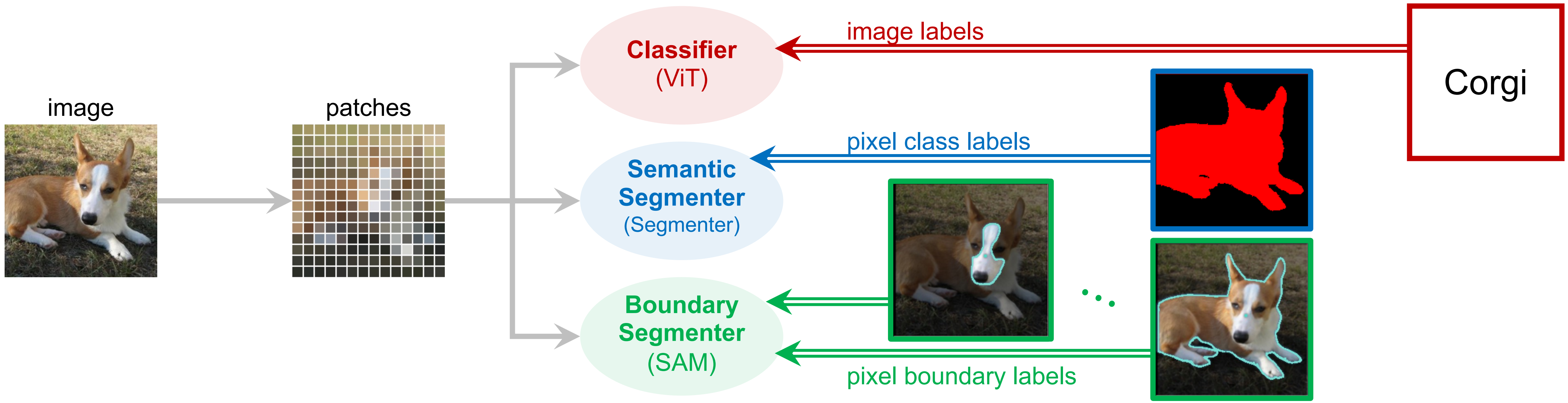
Prior works: Build separate models and supervision



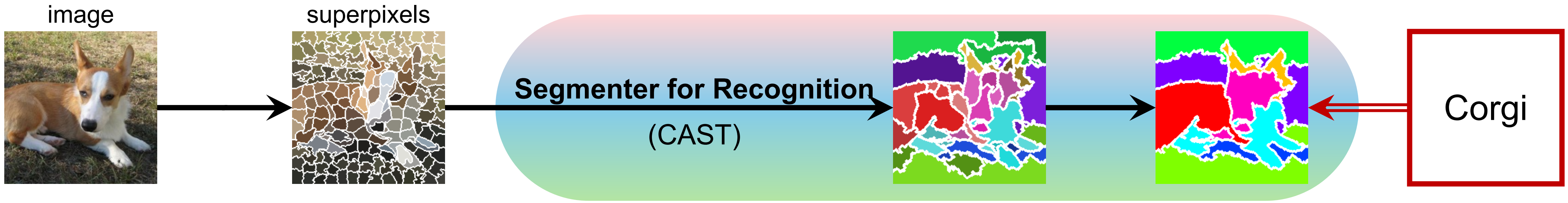
Segment Anything. Kirillov et al. ICCV 2023

Problems of prior works:

1. Need separate annotations
2. Need separate models
3. One task does not help the other

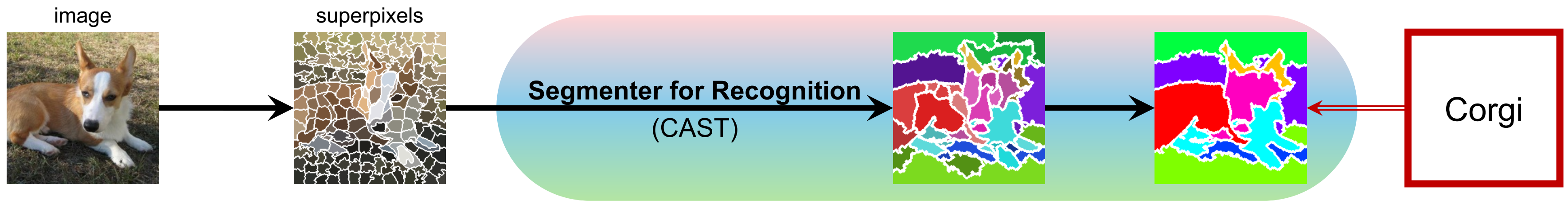


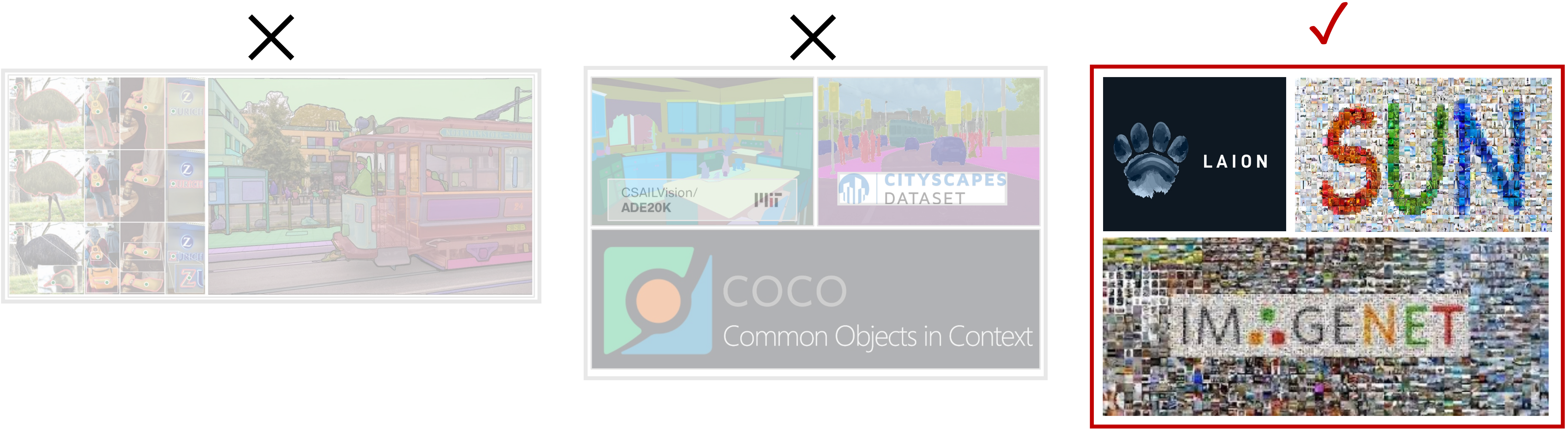
Our idea: put segmentation in the loop of recognition



Strengths of our idea

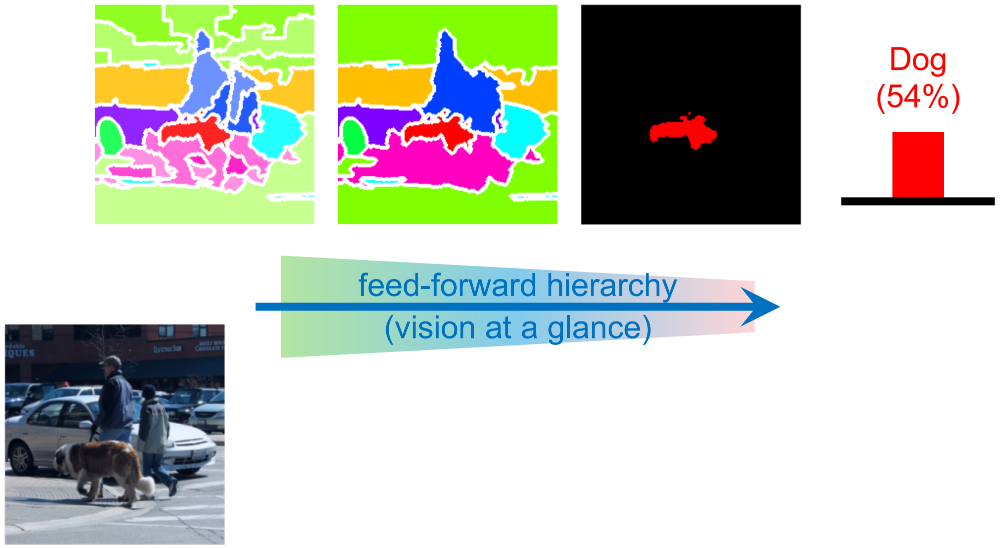
1. Learn hierarchical segmentation FOR FREE from image labels





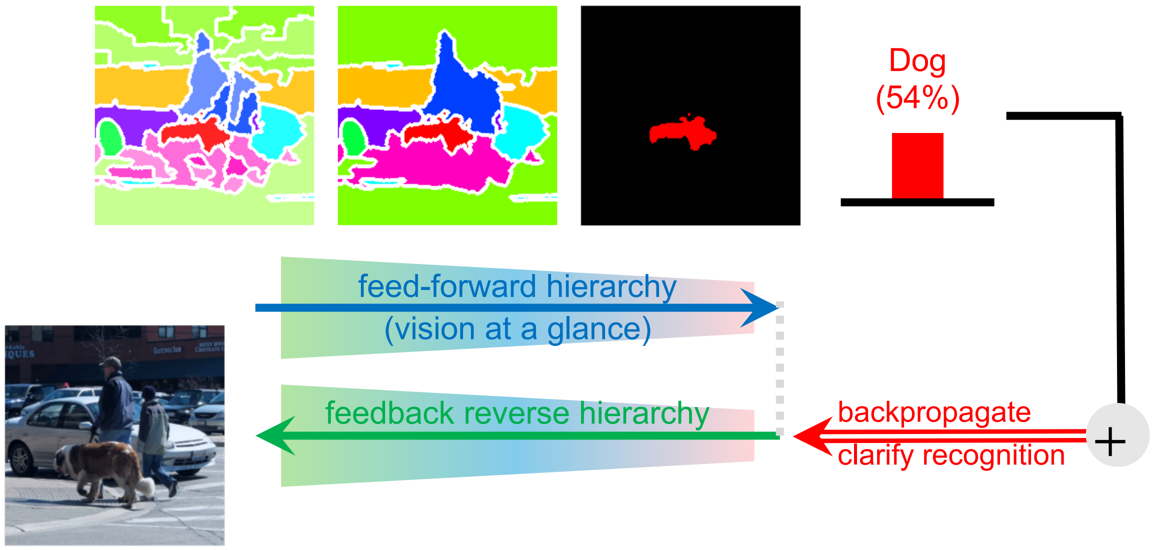
Strengths of our idea

1. Hierarchical segmentation learned FOR FREE from image labels
2. Adaptive segmentation improved with image recognition



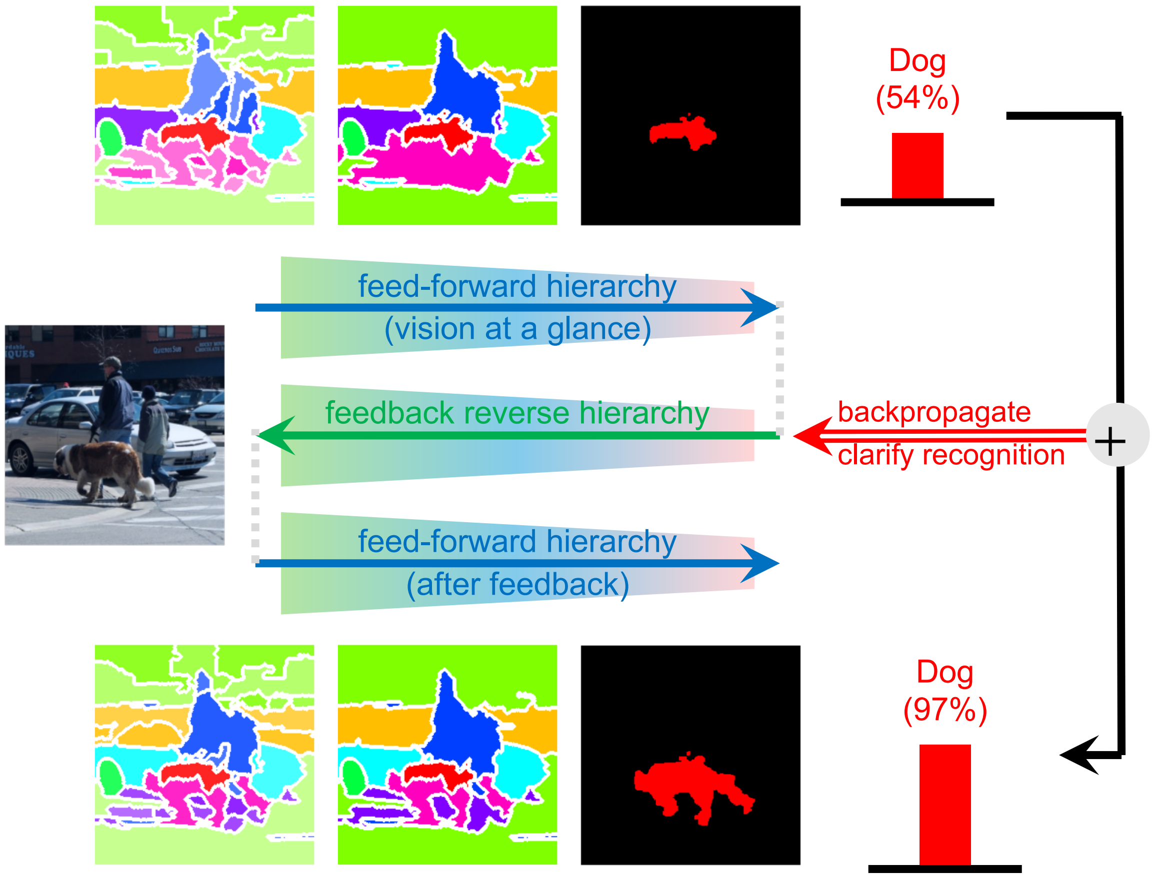
Strengths of our idea

1. Hierarchical segmentation learned FOR FREE from image labels
2. Adaptive segmentation improved with image recognition

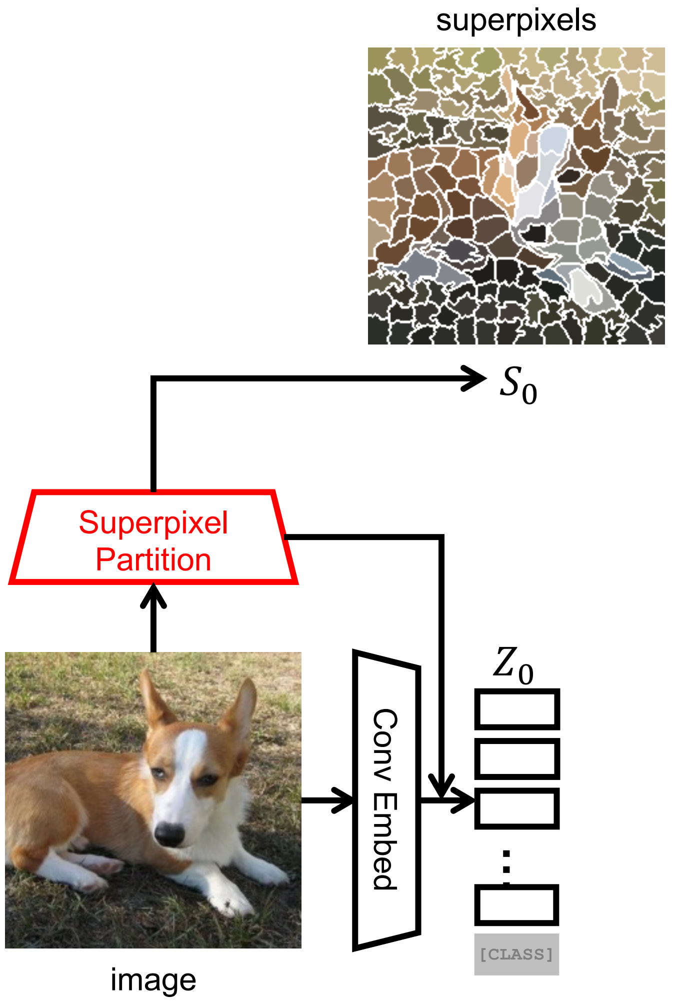


Strengths of our idea

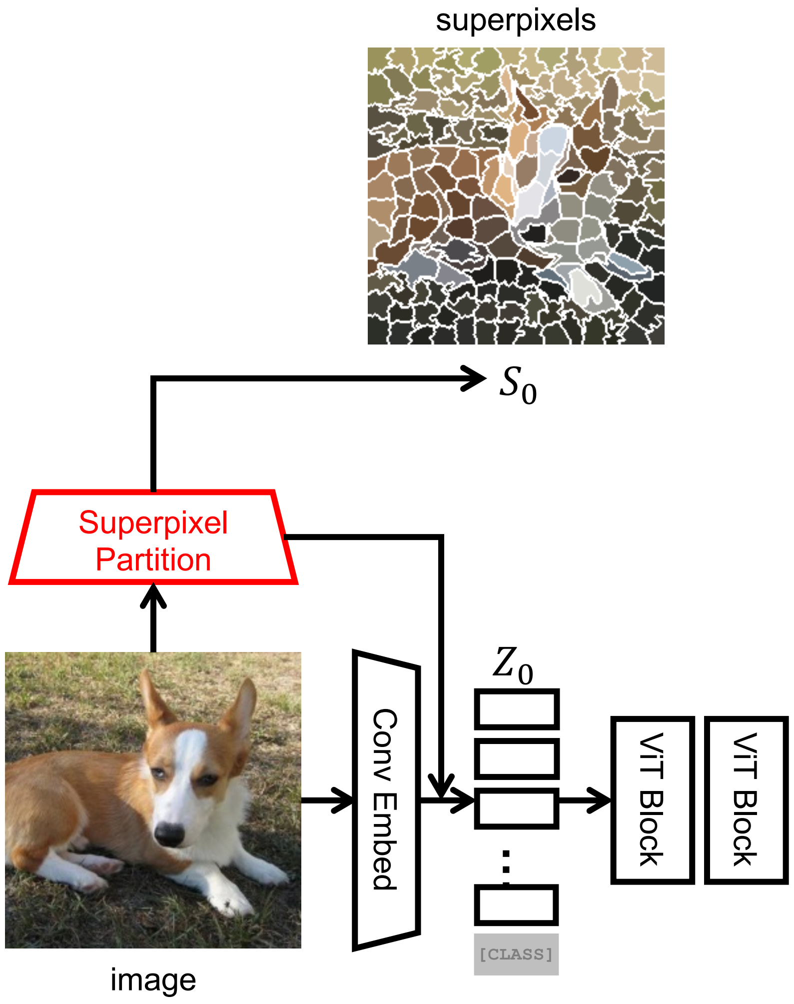
1. Hierarchical segmentation learned FOR FREE from image labels
2. Adaptive segmentation improved with image recognition



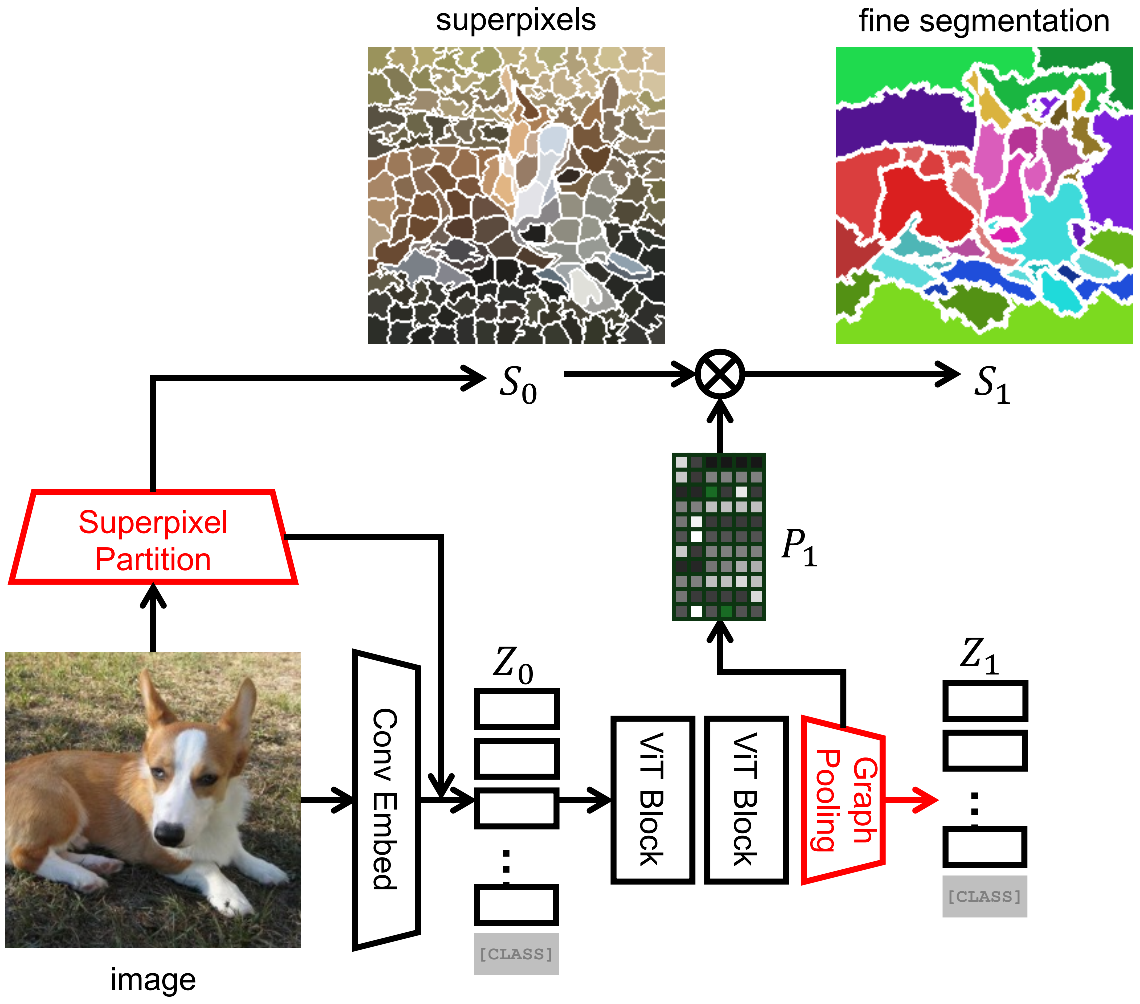
Step 1: begin with segment (superpixel) tokens



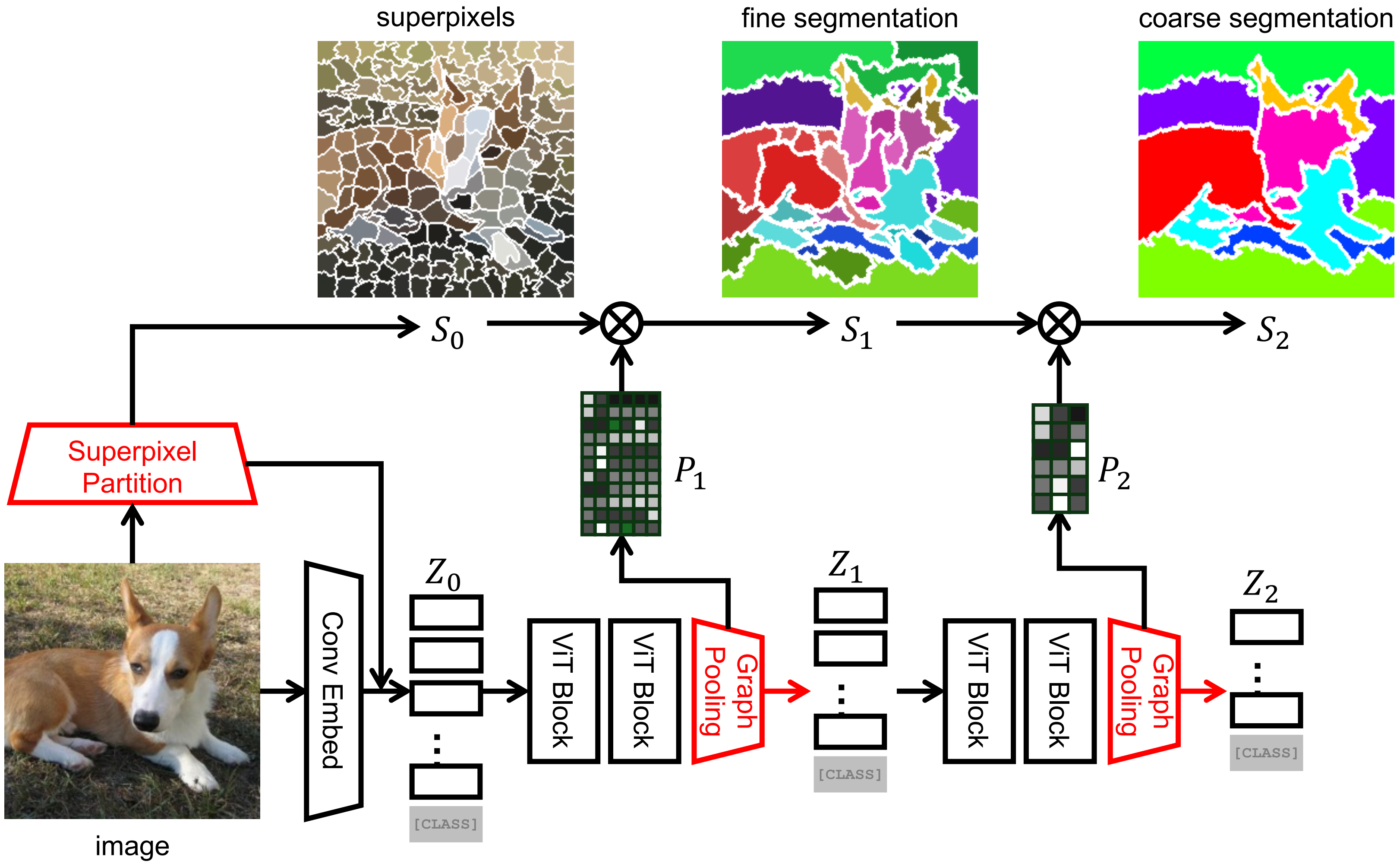
Step 2: contextualize tokens with transformer encoder



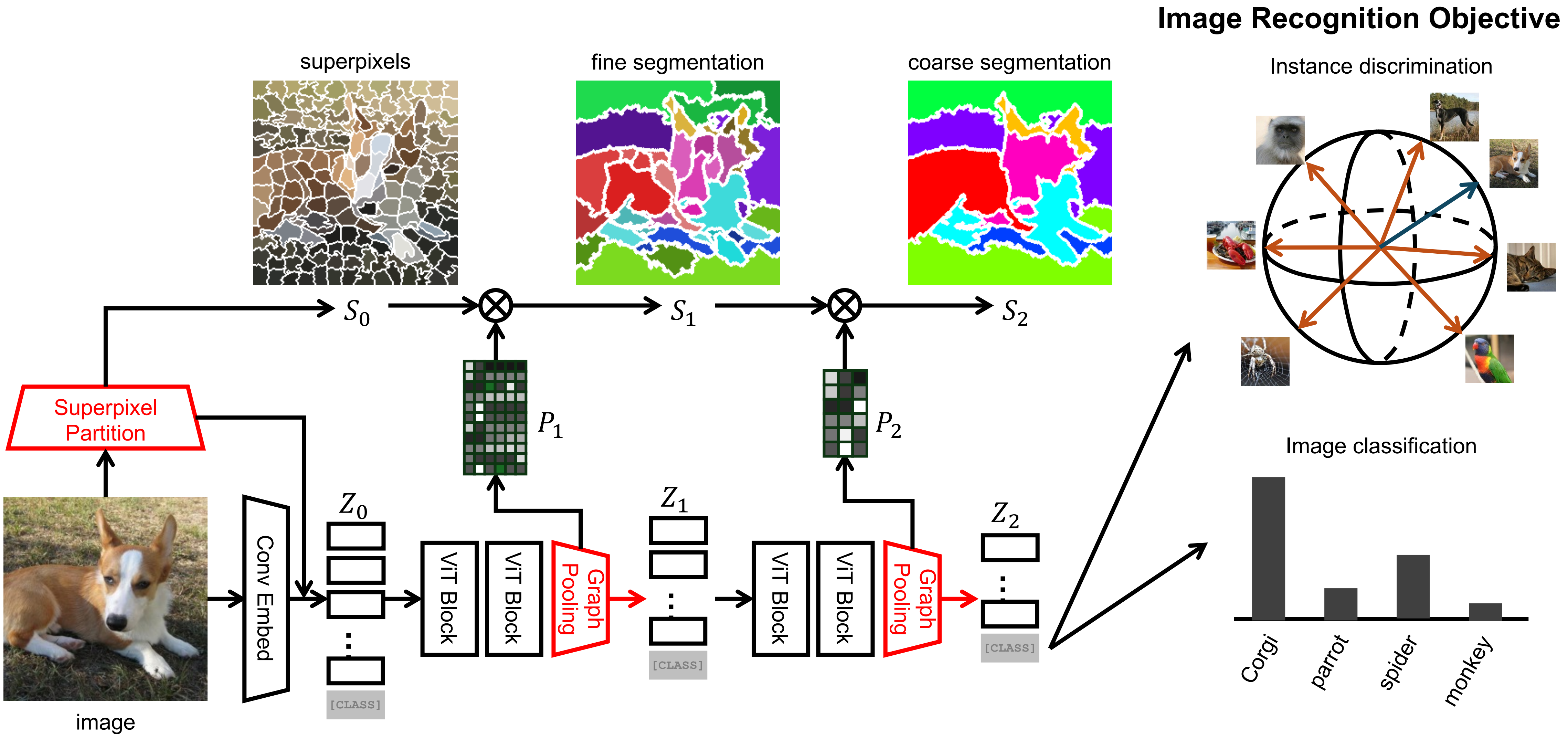
Step 3: group fine segments to coarse regions



Repeat step 2 to 3

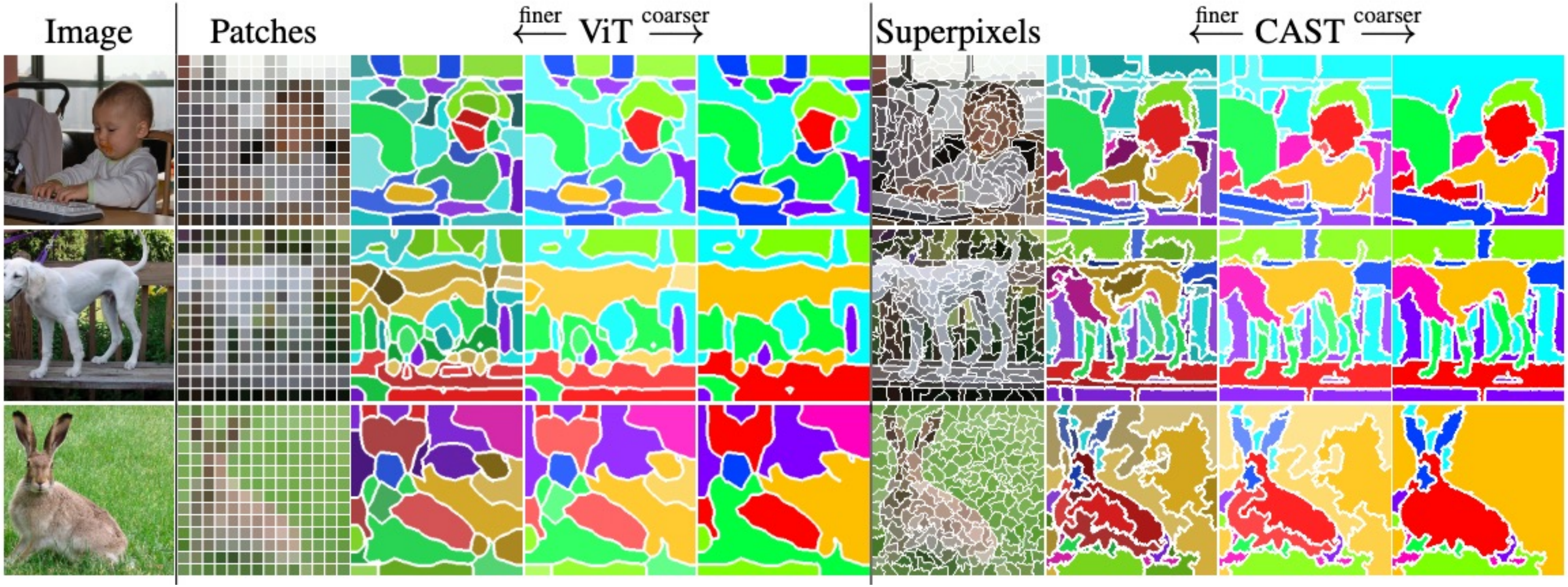


Step 4: predict image-level recognition



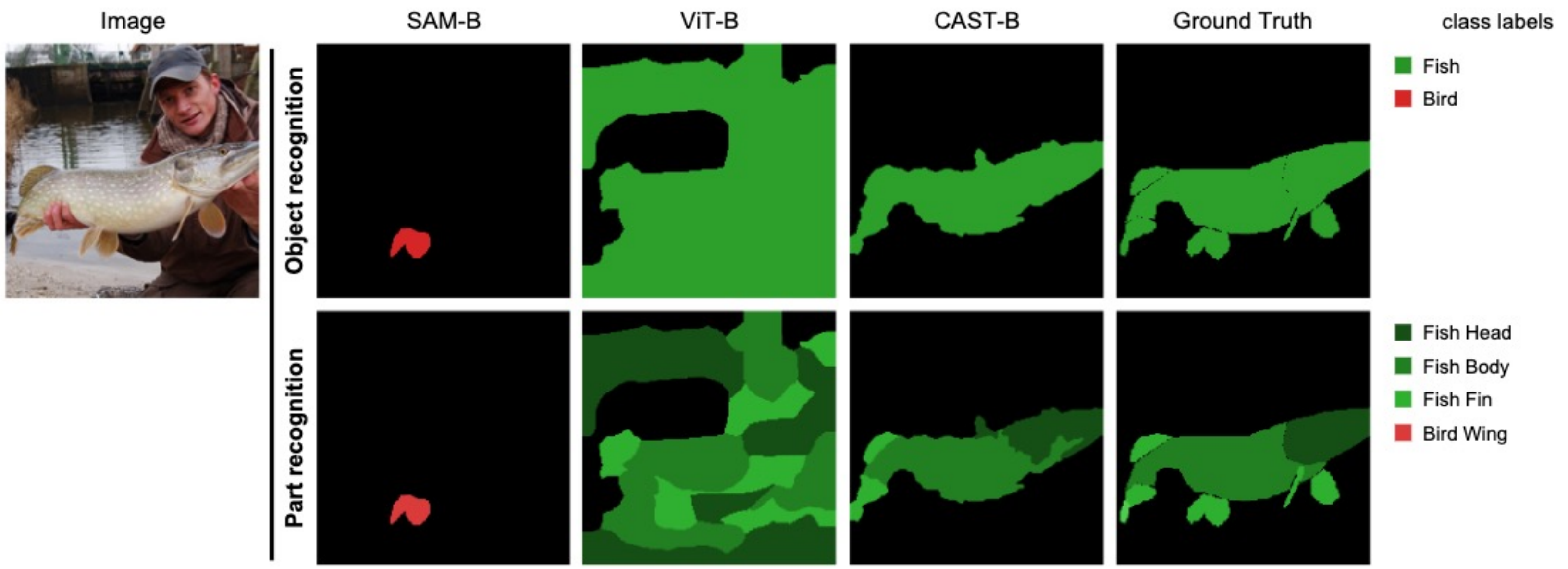
Benefits of CAST

1. Unsupervised discovery of hierarchical segmentations



Benefits of CAST

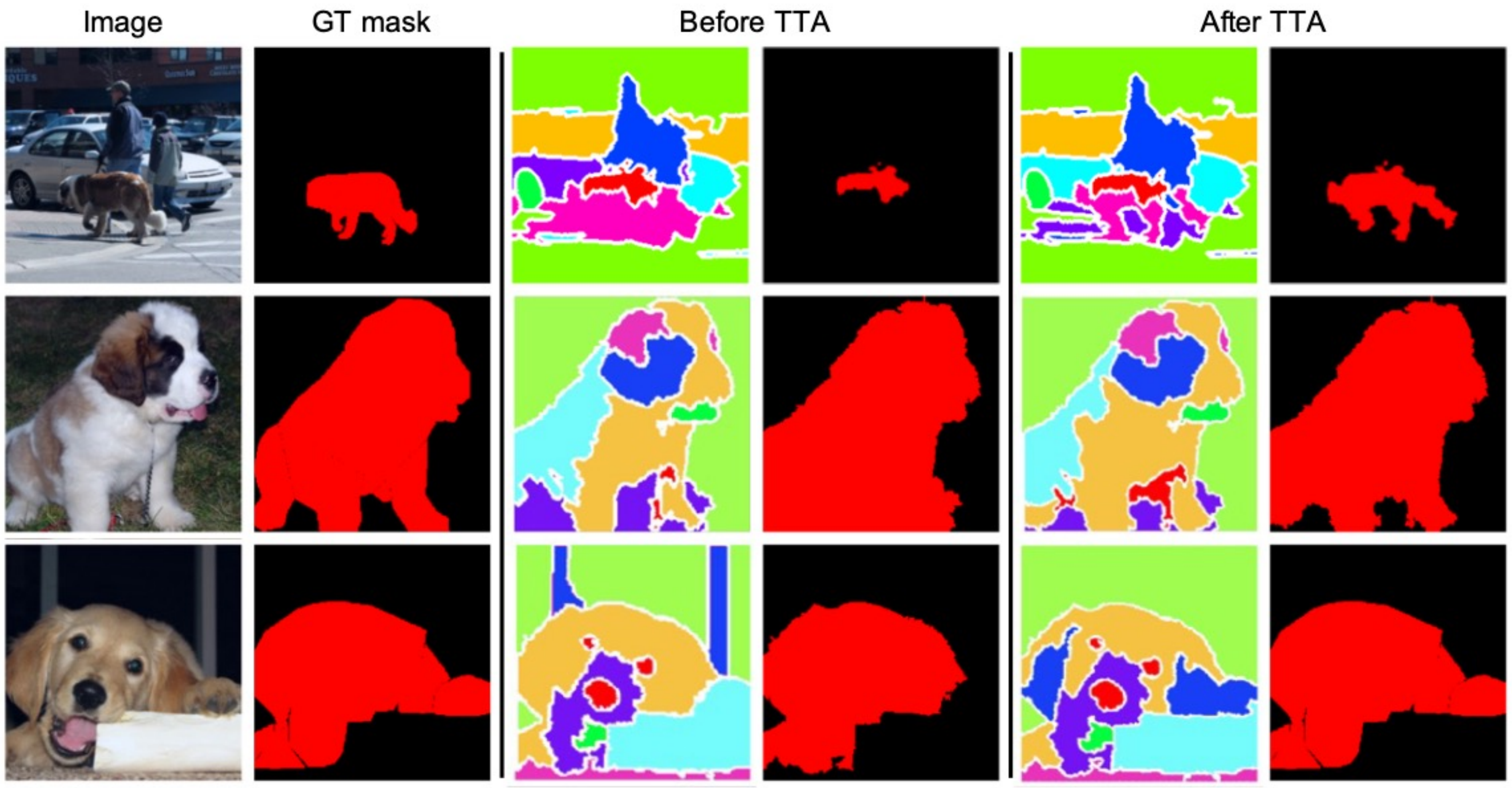
1. Unsupervised discovery of hierarchical segmentations
2. Unsupervised part / object segmentation beats Segment Anything



Model	Training data	Supervised	GFLOPS	Part $\overset{\text{finer}}{\longleftarrow}$	Object $\overset{\text{coarser}}{\longrightarrow}$	Category
SAM-B	SA-1B	✓	488.2	10.15 / **7.25**	18.03 / 20.71	31.36 / 32.01
ViT-B	IN-1K	✗	17.8	11.74 / 4.64	25.34 / 10.92	36.68 / 13.28
CAST-B	IN-1K	✗	**12.9**	**13.20** / 6.52	**29.66** / **22.32**	**50.75** / **34.38**

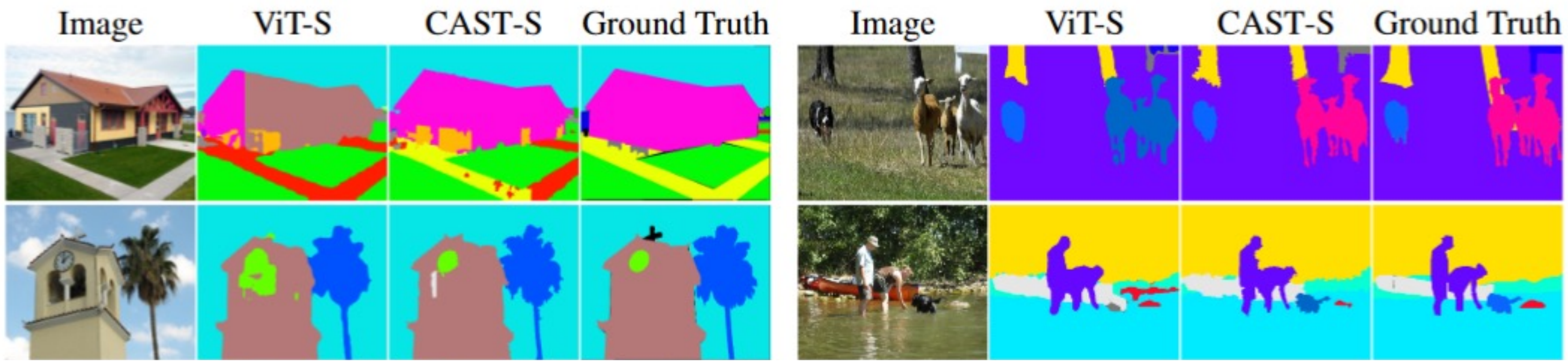
Benefits of CAST

1. Unsupervised discovery of hierarchical segmentations
2. Unsupervised part / object segmentation beats Segment Anything
3. Concurrent segmentation and recognition during inference



Benefits of CAST

1. Unsupervised discovery of hierarchical segmentations
2. Unsupervised part / object segmentation beats Segment Anything
3. Concurrent segmentation and recognition during inference
4. Better performance and efficiency than ViT



segmentation

(a) Pascal VOC	Token	Pooling	Before tuning	After tuning
ViT-S	Patch	✗	30.9 / 16.1	65.8 / 40.7
↕ ablation	Patch	✓	34.5 / 19.8	67.2 / 41.9
	Superpixel	✗	32.2 / 21.2	66.5 / 46.7
CAST-S	Superpixel	✓	**38.4 / 27.0**	**67.6 / 48.1**

classification & efficiency

Model	GFLOPS	IN-100	IN-1K
ViT-S	4.7	78.1	67.9
Swin-T	4.5	78.3	63.0
CAST-S	**3.4**	**79.9**	**68.1**

Thank you for your listening

Code available at:
https://github.com/twke18/CAST