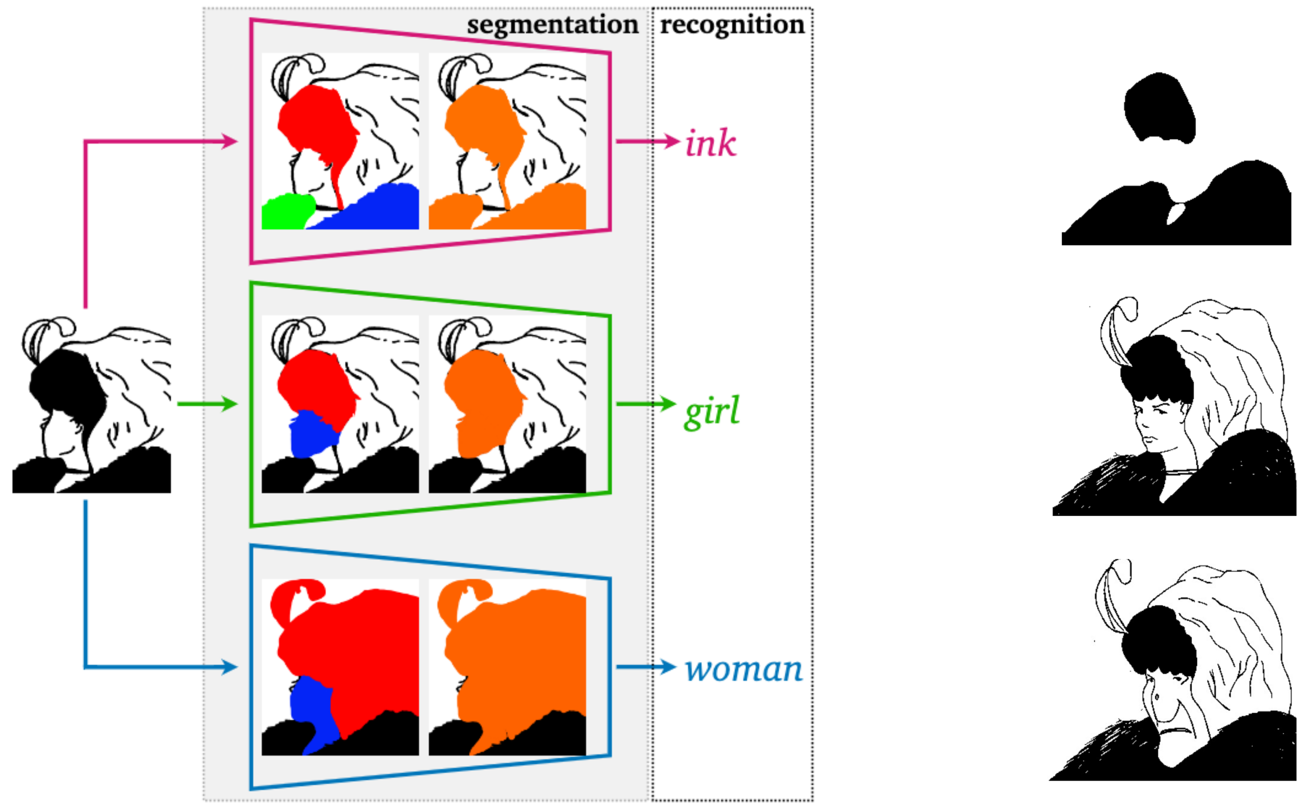
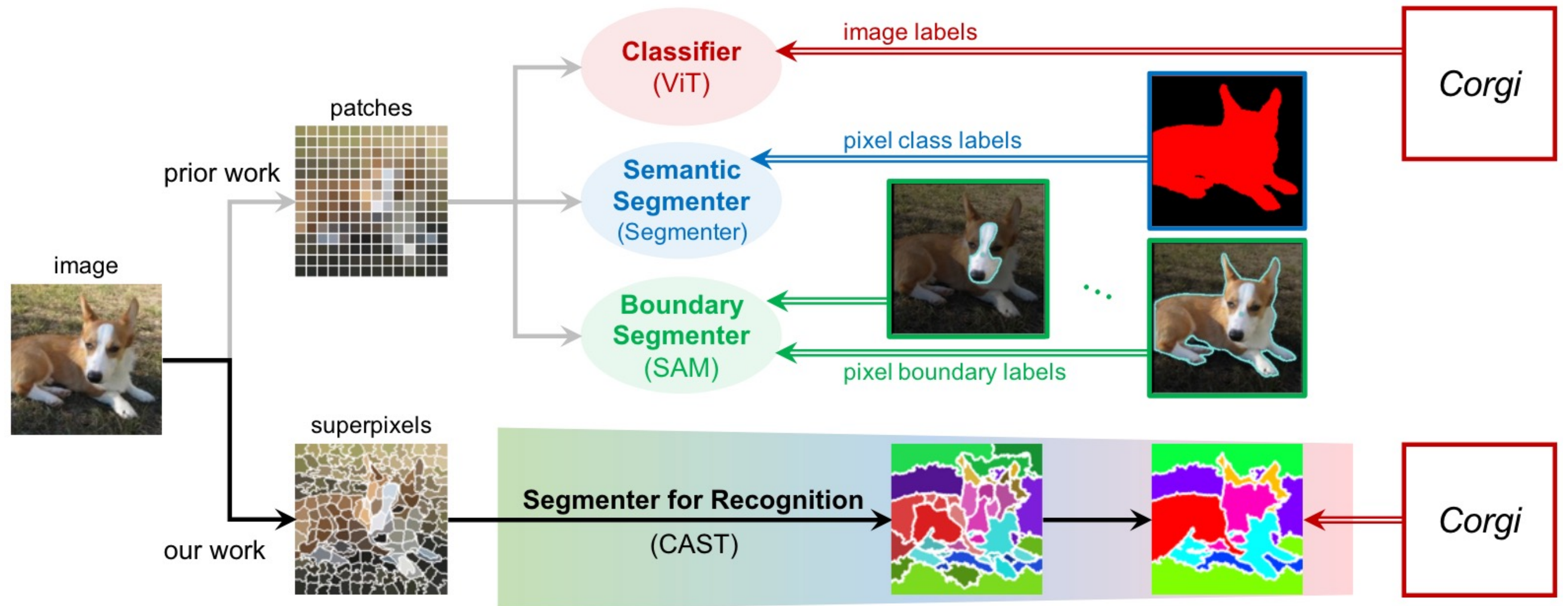
# Learning Hierarchical Image Segmentation For Recognition and By Recognition


Tsung-Wei Ke *   Sangwoo Mo *   Stella X. Yu

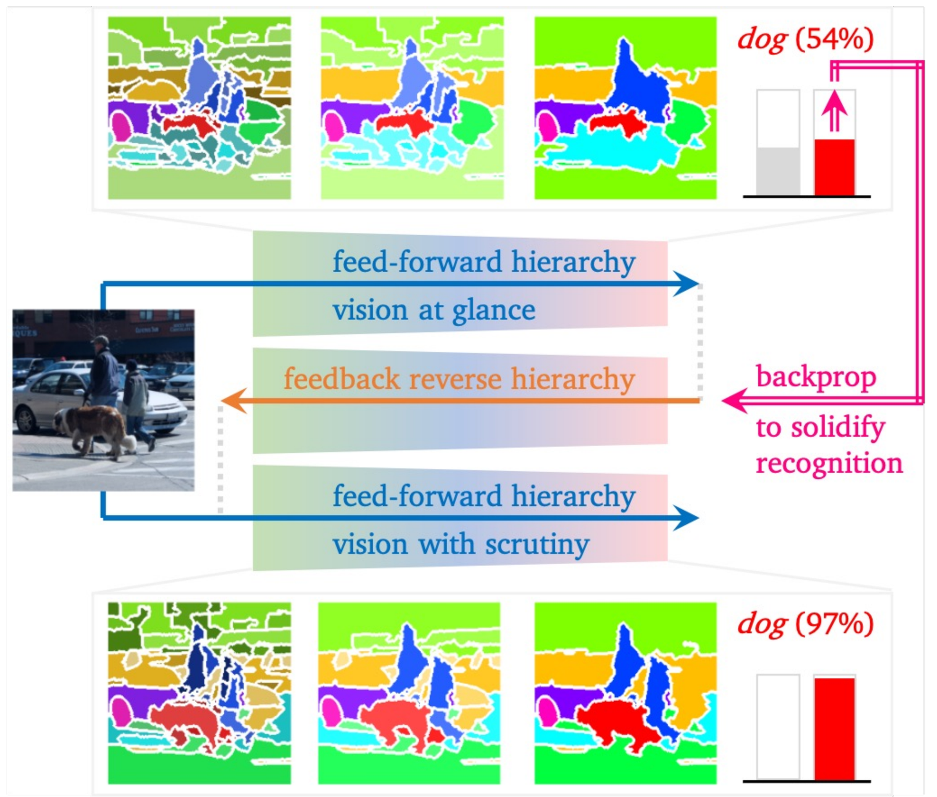

## Consistency of Visual Parsing instead of Text Labels




## Continuum of Internal Segmentation onto Recognition

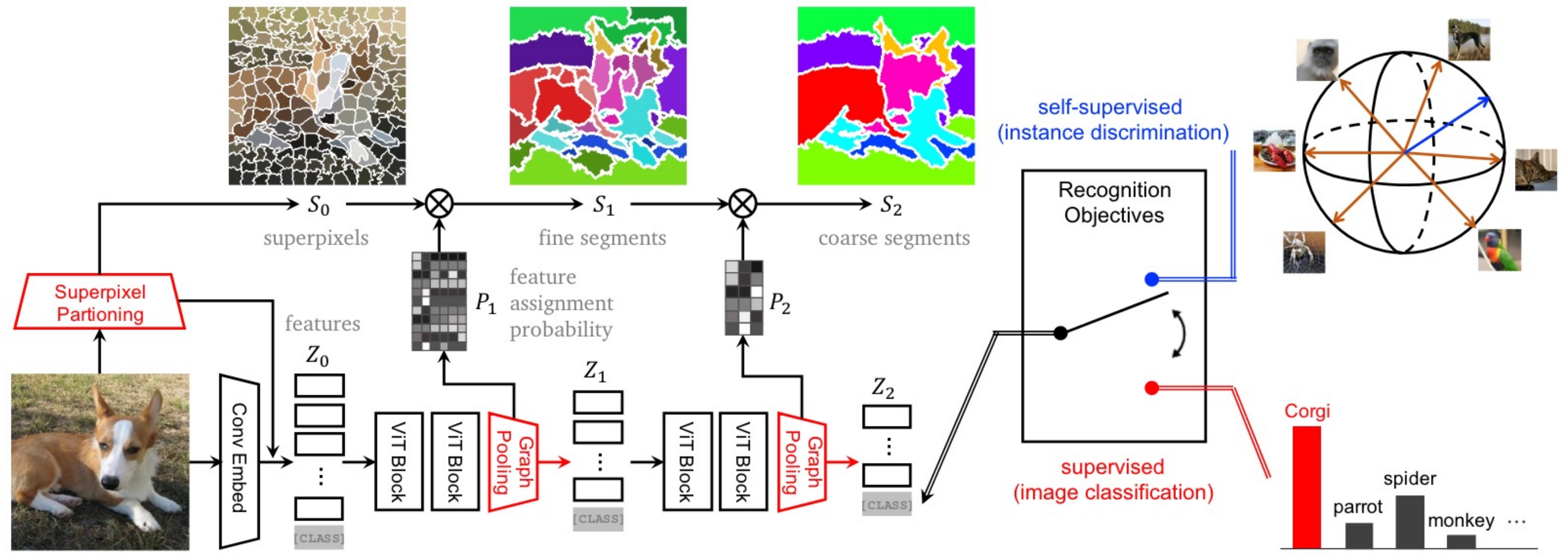



## Concurrency of Segmentation and Recognition



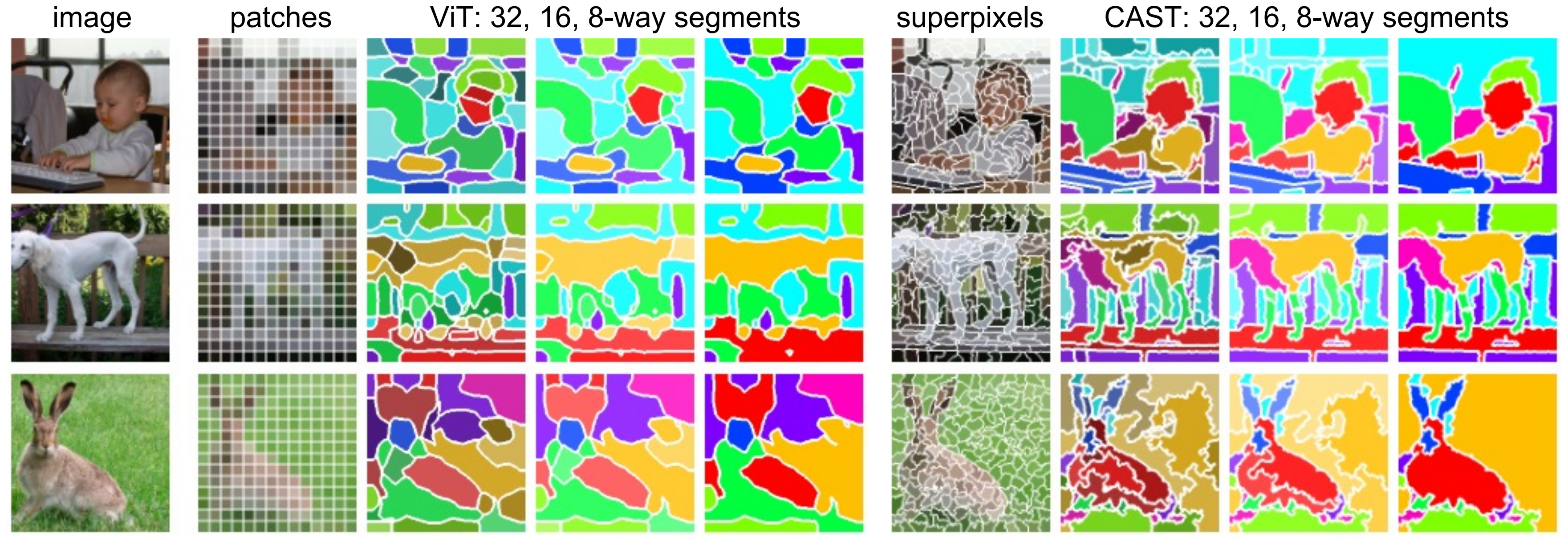

1. Always trained, tested, and adapted together
2. Segmentation substantiates recognition
3. Recognition leads segmentation

## Our CAST = ViT w/ Superpixels + Graph Pooling

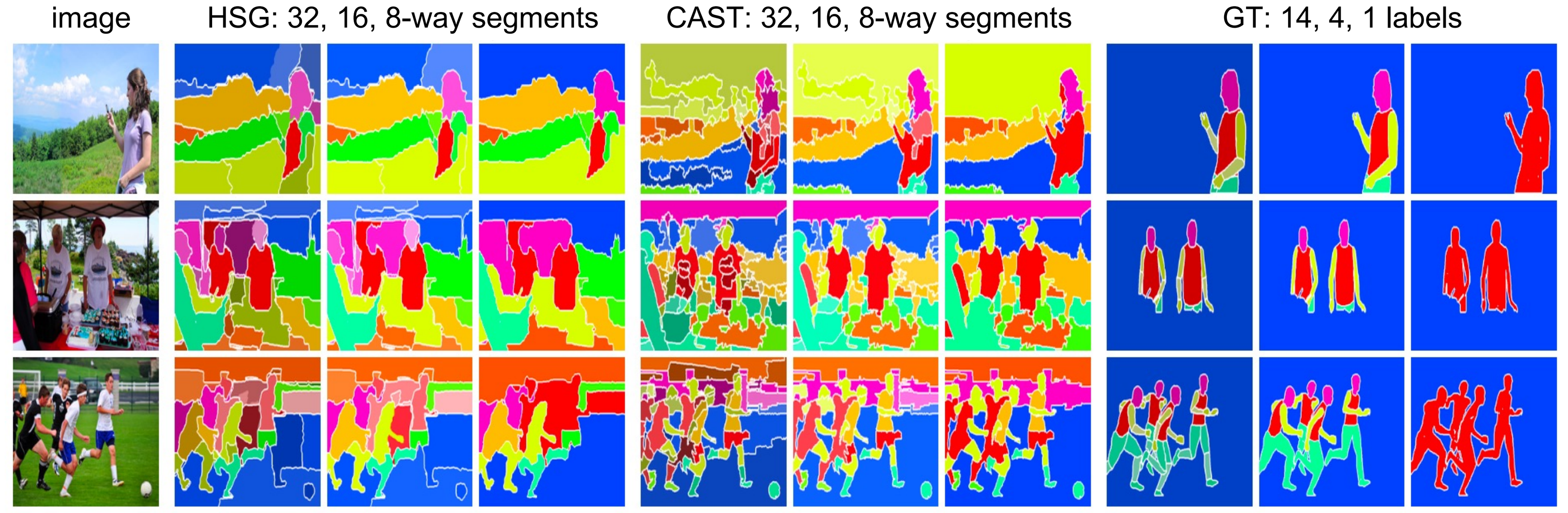



| | | |
|---|---|---|
| **Segmentation** as part | **OF** | **Recognition** |
| learned | **BY** | **Recognition** |
| uncovering part-whole | **FOR** | **Recognition** |

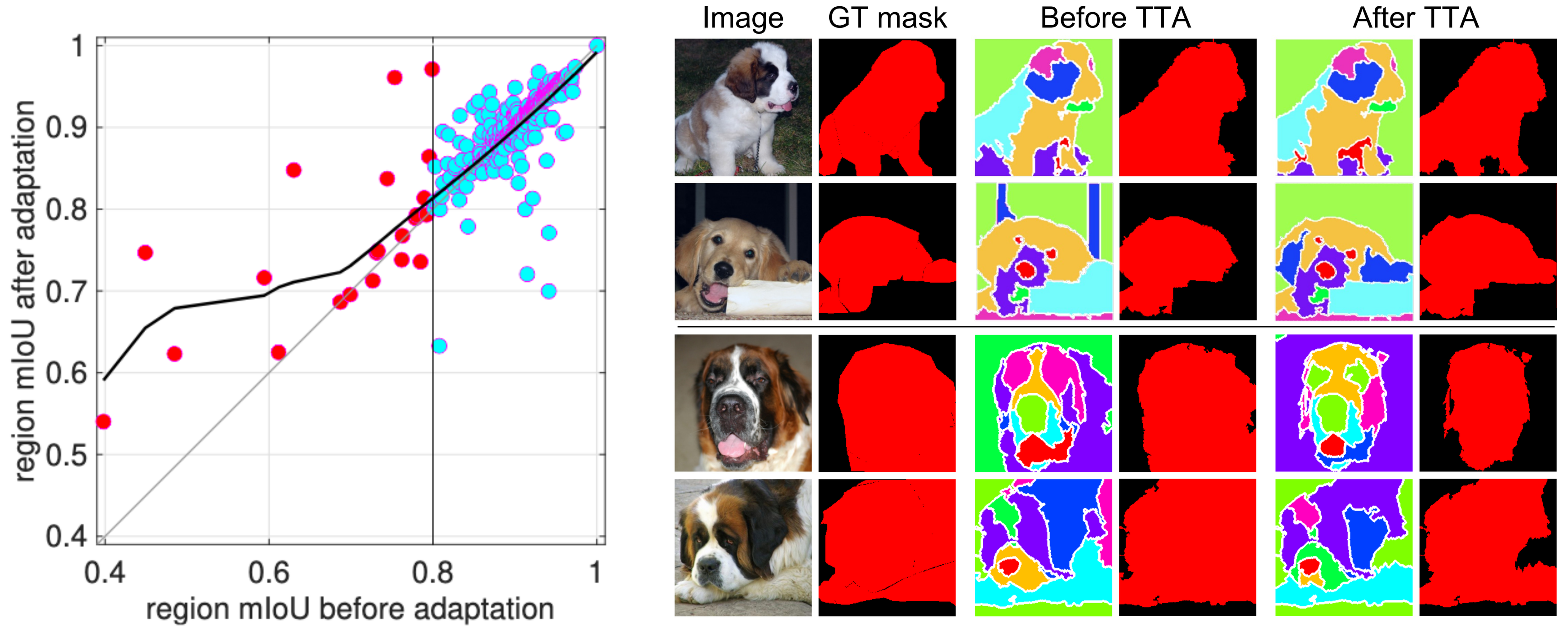
## 1. Unsupervised Part-Whole Discovery



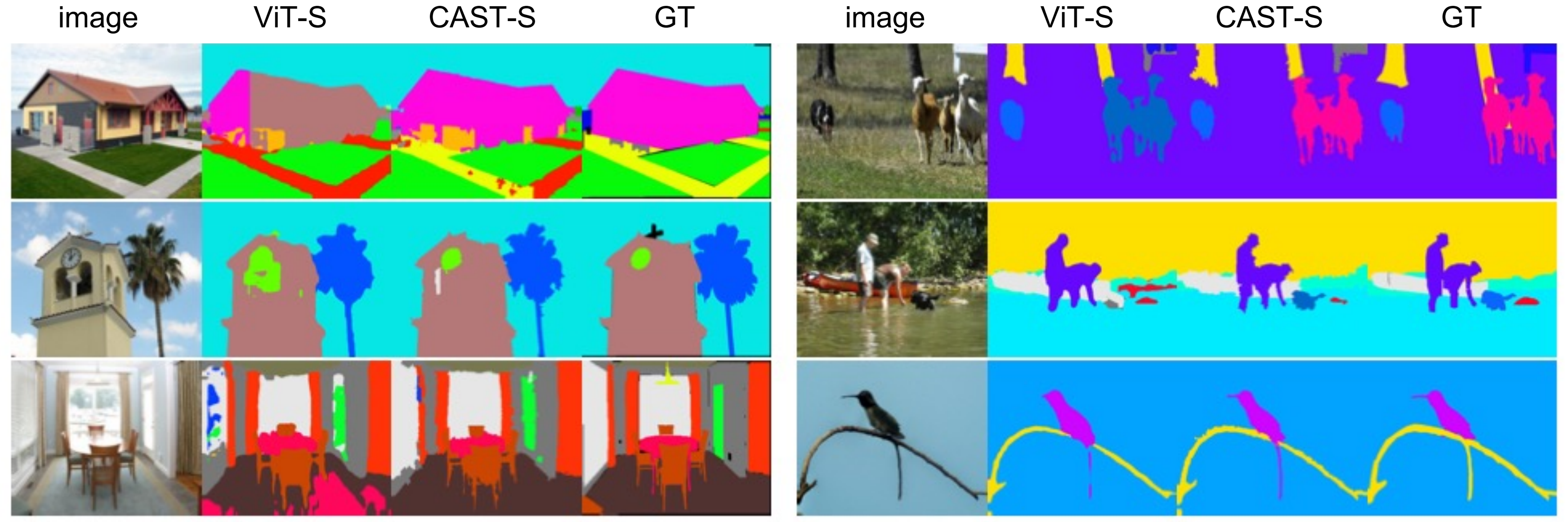





## 2. Solidify Recognition by Adapting Segmentation




## 3. Generalization and Efficiency




| model | GFLOPS | region mIoU | | boundary F-score | |
|---|---|---|---|---|---|
| SAM-B | 677 | 18.03 | 10.15 | 20.71 | 7.25 |
| SAM-H | 3166 | 21.97 | 12.07 | **32.66** | **11.82** |
| ViT-B | 18 | 25.34 | 11.74 | 10.92 | 4.64 |
| CAST-B | **13** | **29.66** | **13.20** | 22.32 | 6.52 |

| P R F | 14 labels | | | 4 labels | | | 1 label | | |
|---|---|---|---|---|---|---|---|---|---|
| HSG | 20.7 | 18.6 | 19.6 | 24.1 | 30.6 | 26.9 | 20.5 | 36.1 | 26.2 |
| CAST | **21.1** | **24.1** | **22.5** | **24.8** | **33.2** | **28.4** | **26.3** | **44.9** | **33.2** |
| gain | 0.4 | 5.5 | 2.9 | 0.7 | 2.6 | 1.5 | 5.8 | 8.8 | 7.0 |

| test on PASCAL-VOC | before tuning | | after tuning | |
|---|---|---|---|---|
| ViT-S | 30.9 | 16.1 | 65.8 | 40.7 |
| ViT-S but with token pooling | 34.5 | 19.8 | 67.2 | 41.9 |
| ViT-S but with superpixels | 32.2 | 21.2 | 66.5 | 46.7 |
| CAST-S | **38.4** | **27.0** | **67.6** | **48.1** |

| 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** |

1. Beats ViT on unsupervised hierarchical segmentation on ImageNet
2. Beats SAM on unsupervised object segmentation on PartImageNet
3. Beats HSG on unsupervised human-body parsing on DensePose
4. Better downstream semantic segmentation
5. Better recognition and efficiency
6. Both superpixels and token pooling contribute to performance gains