Solving the task of semantic segmentation requires identifying objects of different sizes and shapes which require the model to extract features at different granularities. Feature Pyramid Network-like methods allow to extract multiple feature representations at different scales with a single forward. What is the optimal to use these multi-scale features for semantic segmentation?

**Introduction**

Merging strategies

- **mIoU:** averaging 44.4
- **product:** majority vote 39.89
- **hierarchical att.** explicit att. 39.7

Self-Ensemble?

- **d1, d2, d3, d4:** pascal, coco, city

| Method | mIoU | #params. | FLOPs |
|--------|------|---------|-------|
| UperNet | 42.02 | 67M | 238G |
| SenUperNet | 42.8 | 70M | 135G |
| FFBaseline | 43.1 | 52M | 307G |
| SenFormer | 44.3 | 55M | 179G |

**Self-Ensemble**

- a) One learner per scale
- b) Merge the predictions
- c) Weight Sharing for parameters mitigation

**Learners architecture**

**Merging strategies**

- **Weight sharing architecture**

- **P:** features from the FPN
- **cls:** learnable embeddings (one per class in the dataset)

**Results**

**A2E0K**

- **Method**
  - DeepLabV3+ [5] R50 44.0 44.9
  - PerPixBaseline [6] R50 41.9 42.9
  - MaskFormer [6] R50 44.5 46.7
  - SenFormer [7] R50 47.2 49.2
  - OCRNet [46] R101 45.5 47.2
  - DeepLabV3+ [5] R101 45.5 47.2
  - SenFormer [7] R101 46.9 47.9
  - OCRNet [46] R101 53.2 54.3

**COCO-Stuff-10K**

- **Method**
  - DANet [11] R50 50.5
  - EMANet [26] R50 50.5
  - CAA [21] R50 50.2
  - SenFormer [7] R50 53.18 54.3
  - DANet [11] R101 50.5
  - EMANet [26] R101 50.5
  - CAA [21] R101 50.2
  - SenFormer [7] R101 55.0

**Pascal-Context**

- **Method**
  - EMA Net [16] R50 37.6
  - MaskFormer [6] R50 38.1 39.8
  - SENetFormer [7] R101 41.0 40.6
  - OCRNet [32] R101 40.5
  - CAA [14] EN-B7 - 45.4
  - SenFormer [7] Swin-L 49.8 51.5

**SenFormer vs UperNet**

- **Method**
  - UperNet ResNet-50 512 × 512 67M 238G 42.05
  - SenFormer ResNet-50 512 × 512 55M 179G 44.38
  - UperNet ResNet-101 512 × 512 86M 257G 43.82
  - SenFormer ResNet-101 512 × 512 79M 199G 46.93
  - UperNet Swin-T 512 × 512 60M 236G 44.41
  - SenFormer Swin-T 512 × 512 59M 179G 46.0
  - UperNet Swin-S 512 × 512 81M 259G 47.72
  - SenFormer Swin-S 512 × 81M 208G 49.2
  - UperNet Swin-B1 640 × 640 121M 471G 50.04
  - SenFormer Swin-B1 640 × 640 120M 371G 52.21
  - UperNet Swin-L1 640 × 640 234M 647G 52.05
  - SenFormer Swin-L1 640 × 640 233M 546G 53.08

**Introduction**

Efficient Self-Ensemble for Semantic Segmentation

Walid Bousselham, Guillaume Thibault, Lucas Pagano, Archana Machireddy, Joe Gray, Young Hwan Chang, Xubo Song