Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis and Luc Van Gool
Towards Unsupervised Semantic Segmentation

**Problem:** How to learn dense semantic representations without supervision?

→ Most works rely on annotations:
  - Weakly supervised: scribbles, bounding boxes, tags
  - Semi supervised: fraction is labeled

→ **Our focus:** learn pixel-level representations for semantic segmentation without using ground-truth

Obukhov et al., “Gated CRF loss for weakly supervised semantic image segmentation” [Figure]
Prior work – Three paradigms

I. Representation Learning

_Idea:_ (1) Solve a pretext task to learn meaningful representations without annotations + (2) offline clustering

**Image-level:**
- Ex: instance discrimination
- Image based
- Background can dominate

**Patch-level:**
- Ex: Colorization
- Proxy task is not decoupled (covariant)

**Limitations:**
- Small-scale datasets with narrow visual domain
- Cluster learning latches onto low-level features
- Special mechanisms required (Sobel filtering)

II. End-To-End Learning

_Idea:_ - Maximize mutual information between an image and its augmentations at pixel level

**Limitations:** - Small-scale datasets with narrow visual domain
- Cluster learning latches onto low-level features
- Special mechanisms required (Sobel filtering)

III. Boundary supervision

_Idea:_ - Obtain semantic segments from boundaries

**Limitations:** - Annotated boundaries
- K-Means?

[1] Ji et al., Invariant information clustering for unsupervised image classification and segmentation. ICCV, 2019.
[2] Larsson et al., Colorization as a proxy task for visual understanding. CVPR, 2017.
[3] Wu et al., Unsupervised feature learning via non-parametric instance discrimination. CVPR, 2018.
Approach (Overview)

Divide-and-conquer strategy:

**Step 1:** Look for regions that likely belong together
→ Shared pixel ownership assumption
→ Use a mid-level visual prior

**Step 2:** Generate semantic pixel embeddings
→ Leverage object mask proposals
→ Maximize or minimize the agreement

Advantages:
- Reduced dependence on the network initialization
- Proxy task is decoupled from feature learning
- Kmeans can be applied to obtain semantics
→ hypothesis: this a more reliable pixel grouping strategy
Perceptual Priors for Grouping Pixels

Criteria:
• No reliance on external supervision
• Strong generalization to new scenes → bottom-up approach

(1) Low-level Vision:
• Handcrafted kernels: intensity, distance, color, texture,...
• Edges or superpixels

(2) Mid-level Vision: → More semantically meaningful
• Saliency:
  - ensemble of handcrafted priors
  - background connectivity, hard edges, Guassian, etc.
• Self-supervised depth / optical flow
MaskContrast: Contrasting Salient Object Masks

**Pixel-Level Objective function:**

\[
\mathcal{L} = -\log \frac{\exp(\Psi_\eta(X)^T \cdot \Psi_\eta(X^+)/\tau)}{\sum_{k=0}^{K} \exp(\Psi_\eta(X)^T \cdot \Psi_\eta(X^-_k)/\tau)}
\]

\[
\mathcal{L}_i = -\log \frac{\exp (z_i \cdot z_{M_{X+}}/\tau)}{\sum_{k=0}^{K} \exp (z_i \cdot z_{M_{X^-_k}}/\tau)}
\]

\[
z_{M_n} = \frac{1}{|M_n|} \sum_{i \in M_n} z_i
\]

Mined masks = \{M_0, M_1, \ldots, M_N\}

Positive pairs = (z_i, z_{M_{X+}}) for i ∈ M_X

Negative pairs = (z_i, z_{M_{X^-_k}})

- **Pull force:** Maximize the agreement between pixels belonging to the same (augmented) mask.
- **Push force:** Avoid mode collapse in the embedding space by driving pixels from different masks apart.
I. Experiments: Setup and Ablations

Training setup:
- Unsupervised Saliency\(^1\) / supervised saliency\(^2\)
- DeeplabV3 (dilated ResNet50)
- Similar to MoCo’s setup (augmentation + memory bank + momentum)

Ablations (PASCAL VOC):

| Mask Proposals             | LC (MIoU) | Augmented Views | Memory | Momentum Encoder | LC (MIoU) |
|----------------------------|-----------|-----------------|--------|------------------|-----------|
| Hierarchical Seg.          | 30.5      | X               | X      | X                | 52.4      |
| Unsupervised Sal. Model    | 58.4      | ✓               | X      | X                | 54.0      |
| Supervised Sal. Model      | 62.2      | ✓               | ✓      | ✓                | 55.0      |

(a) Comparison of three mask proposal mechanisms.

(b) Analysis of the used training mechanisms.

- Regions extracted with the hierarchical segmentation algorithm were often too small to be representative of an object or part.
- Mid-level visual prior is beneficial.

\(^1\) Nguyen et al., *Deepusps: Deep robust unsupervised saliency prediction via self-supervision*. NeurIPS, 2019.

\(^2\) Qin et al., *Basnet: Boundary-aware salient object*. CVPR, 2019.
II. Experiments: Linear Classifier and Clustering (PASCAL)

| Method                           | LC   | K-Means |
|----------------------------------|------|---------|
| **Proxy task based:**            |      |         |
| Co-Occurrence                    | 13.5 | 4.0     |
| CMP                              | 16.5 | 4.3     |
| Colorization                     | 25.5 | 4.9     |
| **Clustering based:**            |      |         |
| IIC                              | 28.0 | 9.8     |
| **Contrastive learning based:**  |      |         |
| Inst. Discr.                     | 26.8 | 4.4     |
| MoCo v2                          | 45.0 | 4.3     |
| InfoMin                          | 45.2 | 3.7     |
| SWAV                             | 50.7 | 4.4     |
| **Boundary based:**              |      |         |
| SegSort†                         | 36.2 | -       |
| Hierarch. Group.†                | 48.8 | -       |
| ImageNet (IN) Classifier (Supervised) | 53.1 | 4.7     |
| MaskContrast (MoCo Init. + Unsup. Sal.) | 58.4 | 35.0     |
| MaskContrast (MoCo Init. + Sup. Sal.) | 62.2 | 38.9     |
| MaskContrast (IN Sup. Init. + Unsup. Sal.) | 61.0 | 41.6     |
| MaskContrast (IN Sup. Init. + Sup. Sal.) | **63.9** | **44.2** |

**MaskContrast:**

→ **decouples** feature learning from clustering;

→ is not strongly dependent on the **network initialization**;

→ is more predictive of the semantic segmentation task as we defined a contrastive learning objective at the **pixel-level**;

→ contains **higher-level visual information** compared to the regions obtained from boundary detectors;

→ can be combined with **K-Means** to obtain semantically meaningful clusters.
III. Experiments: Semantic Segment Retrieval (PASCAL)

- Retrieve neighbors from train set for val set
- Evaluate for 7 classes and 21 classes on PASCAL

| Method                        | MIoU (7 classes) | MIoU (21 classes) |
|-------------------------------|------------------|-------------------|
| SegSort                       | 10.2             | -                 |
| Hierarch. Group.              | 24.6             | -                 |
| MoCo v2                       | 48.0             | 39.0              |
| MaskContrast (Unsup. Sal.)    | 53.4             | 43.3              |
| MaskContrast (Sup. Sal.)      | **62.3**         | **49.6**          |

Pascal-S dataset

Query

Nearest neighbors
IV. Experiments: Transfer Learning and Semi-Sup. Learning

**Transfer learning**: PASCAL, COCO and DAVIS datasets (MoCo init.)

| Model                          | PASCAL (MIOU)↑ | COCO (MIOU)↑ | DAVIS ’16 Jm ↑ | Jm ↑ |
|-------------------------------|----------------|--------------|----------------|------|
| MoCo v2                       | 45.0           | 35.2         | 77.1           | 77.2 |
| MaskContrast (Unsup. Sal.)    | 55.4           | 45.0         | 78.0           | 77.8 |
| MaskContrast (Sup. Sal.)      | 57.2           | 47.2         | 82.0           | 80.9 |

**Semi-supervised finetuning** on PASCAL (ImageNet init.)

| Label Fraction | 1%  | 2%  | 5%  | 12.5% | 100% |
|----------------|-----|-----|-----|-------|------|
| ImageNet Classifier Init.    | 43.4| 55.2| 62.7| 68.4  | 78.0 |
| + MaskContrast (Unsup. Sal.) | 50.5| 57.2| 64.5| 69.0  | 78.4 |
| + MaskContrast (Sup. Sal.)   | **51.5**| **59.6**| **65.3**| **69.4**| **78.6** |

Qualitative results with 1% labeled (~100 images)
Qualitative Results (Linear Classifier on PASCAL)
Conclusion

• MaskContrast consists of 2 steps:
  o (1) mine object mask proposals (saliency)
  o (2) learn semantic pixel embeddings through a contrastive loss
• The perceptual prior prevents the model from latching onto low-level image features
• Encouraging clustering results on PASCAL and transfer results to ImageNet/COCO/DAVIS

Future Work

• Extract multiple and more detailed masks for each image
• Use extra sensory data

Code is available on Github

github.com/wvangansbeke/Unsupervised-Semantic-Segmentation