Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers

Tsung-Wei Ke
Jyh-Jing Hwang
Yunhui Guo
Xudong Wang
Stella X. Yu

Berkeley
University of California

CVPR 2022
New Orleans, Louisiana
Task of Semantic Segmentation:
Put Pixels into Semantic Categories

Images \[\rightarrow\] Predictions \[\rightarrow\] Pixel-wise Annotations

- Bike
- Person
- Horse
- Cow
Task of Unsupervised Semantic Segmentation:
Put Pixels into Groups without Any Labeled Supervision
Two Approaches to Predict Pixel Labels from Groupings

Our Model by Feature Learning: Predict Labels from Retrieved Segments

Images -> Segmentation CNN -> Pixel Features

Query: Sofa
Retrievals: Sofa, Chair, Sofa, Sofa, Chair
Current Feature Learning Methods: Contrast Image-Image vs. Pixel-Segment

Contrast images disregarding visual change

Current Feature Learning Methods: Contrast Image-Image vs. Pixel-Segment

Contrast **images** disregarding visual change

Contrast pixels with **regions** w.r.t low-level visual cues


Semantics Intrinsically Has Multiple Levels of Granularity

Coarse-grained Categories

Outdoor Scene

Background

Foreground

Rock

Road

Person

Hand

Torso

Head

Semantics Intrinsically Has Multiple Levels of Granularity

Semantics Intrinsically Has Multiple Levels of Granularity

Most existing methods:
avoid ambiguity / presume a granularity

or

Background  Hand  Torso  Head
Semantics Intrinsically Has Multiple Levels of Granularity

Most existing methods:
- avoid ambiguity / presume a granularity

Our idea: embrace multiple levels of granularity
Face and Body are Parts of a Whole in the Visual Scene
Babies Appear Different but Have the Same Semantics

[Diagram showing baby images and semantic labels like 'room', 'person', 'dining table', etc.]
Invariance: Multiview Cosegmentation

Ground features by visual appearance and correspondence
Invariance: Multiview Cosegmentation

Regularize features by multi-scale grouping
Multi-scale Grouping: Consistency is Not Guaranteed

Grouping Probability at Level $l$: $P_l(a) = \text{Prob}(G_l = a|x)$
Multi-scale Consistency: Clustering Transformer

Grouping Probability at Level $l$:
\[ P_l(a) = \text{Prob}(G_l = a | x) \]

Transition Probability to Level $l+1$:
\[ C_l^{l+1}(a, b) = \text{Prob}(G_{l+1} = b | G_l = a) \]

Grouping Assignment at Level $l+1$:
\[ P_{l+1} = P_l \times C_l^{l+1} = P_0 \times C_0^1 \times \cdots \times C_l^{l+1} \]
Multi-scale Consistency: Clustering Transformer

Grouping Probability at Level $l$:
$$P_l(a) = \text{Prob}(G_l = a | x)$$

Transition Probability to Level $l+1$:
$$C_{l+1}(a, b) = \text{Prob}(G_{l+1} = b | G_l = a)$$

Grouping Assignment at Level $l+1$:
$$P_{l+1} = P_l \times C_{l+1} = P_0 \times C_0^1 \times \cdots \times C_{l+1}$$
Multi-scale Consistency: Clustering Transformer

Grouping Probability at Level $l$:

$$P_l(a) = \text{Prob}(G_l = a | x)$$

Transition Probability to Level $l+1$:

$$C_{l}^{l+1}(a, b) = \text{Prob}(G_{l+1} = b | G_l = a)$$

Grouping Assignment at Level $l+1$:

$$P_{l+1} = P_l \times C_{l}^{l+1} = P_0 \times C_0^{1} \times \cdots \times C_{l}^{l+1}$$
Multi-scale Consistency: Clustering Transformer
Multi-scale Consistency: Clustering Transformer
Our Hierarchical Segment Grouping Model
Our Hierarchical Segment Grouping Model

Pixel-segment contrast loss:
1. Ground features by visual appearance
2. Enforce correspondence across views

\[ L(f) = \lambda_E L_f(G_e) + \lambda_F \sum_{l \geq 1} L_f(G_l) + \lambda_G L_g \]
Our Hierarchical Segment Grouping Model

Pixel-segment contrast loss:
Regularize features by consistent hierarchy

\[ L(f) = \lambda_E L_f(G_e) + \lambda_F \sum_{l \geq 1} L_f(G_l) + \lambda_G L_g \]
Our Hierarchical Segment Grouping Model

Desire balanced, compact, distinctive clusters

Graph Clustering with Graph Neural Networks. Tsitsulin et al. ArXiv 2020.
1. First Unsupervised Hierarchical Semantic Segmentation

\[
\text{NFCovering}(S' \rightarrow S_{fg}) = \frac{1}{|S_{fg}|} \sum_{R \in S_{fg}} \max_{R' \in S'} \frac{|R \cap R'|}{|R \cup R'|}
\]

VOC: Varying number of segmented regions

Normalized Covering (%)
2. SOTA on Unsupervised Semantic Segmentation

<table>
<thead>
<tr>
<th>Training set</th>
<th>MSCOCO</th>
<th>Cityscapes</th>
<th>KITTI-STEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation set</td>
<td>VOC</td>
<td>Cityscapes</td>
<td>KITTI-STEP</td>
</tr>
<tr>
<td>Method</td>
<td>mIoU</td>
<td>Acc.</td>
<td>mIoU</td>
</tr>
<tr>
<td>Moco [20]</td>
<td>28.1</td>
<td>-</td>
<td>15.3</td>
</tr>
<tr>
<td>DenseCL [60]</td>
<td>35.1</td>
<td>-</td>
<td>12.7</td>
</tr>
<tr>
<td>Revisit [56]</td>
<td>35.1</td>
<td>-</td>
<td>17.1</td>
</tr>
<tr>
<td>SegSort [26]</td>
<td>11.7</td>
<td>75.1</td>
<td>24.6</td>
</tr>
<tr>
<td>Our HSG</td>
<td><strong>41.9</strong></td>
<td><strong>85.7</strong></td>
<td><strong>32.5</strong></td>
</tr>
</tbody>
</table>
3. Unsupervised Visual Context Retrievals across Granularity Levels
Code available at https://github.com/twke18/HSG