Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers

Tsung-Wei Ke

Jyh-Jing Hwang

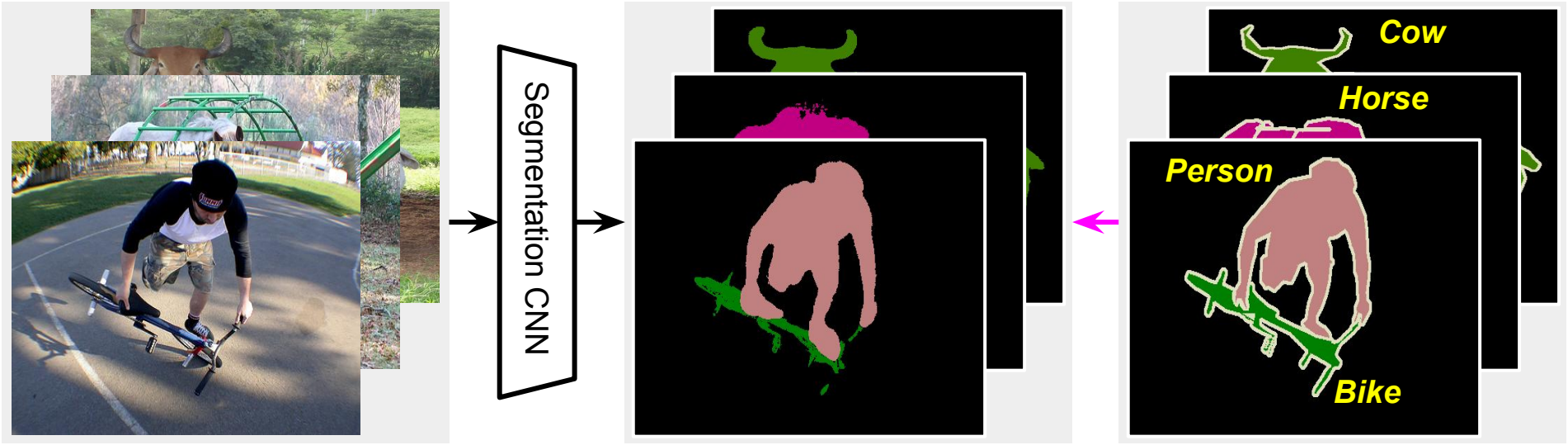
Yunhui Guo

Xudong Wang

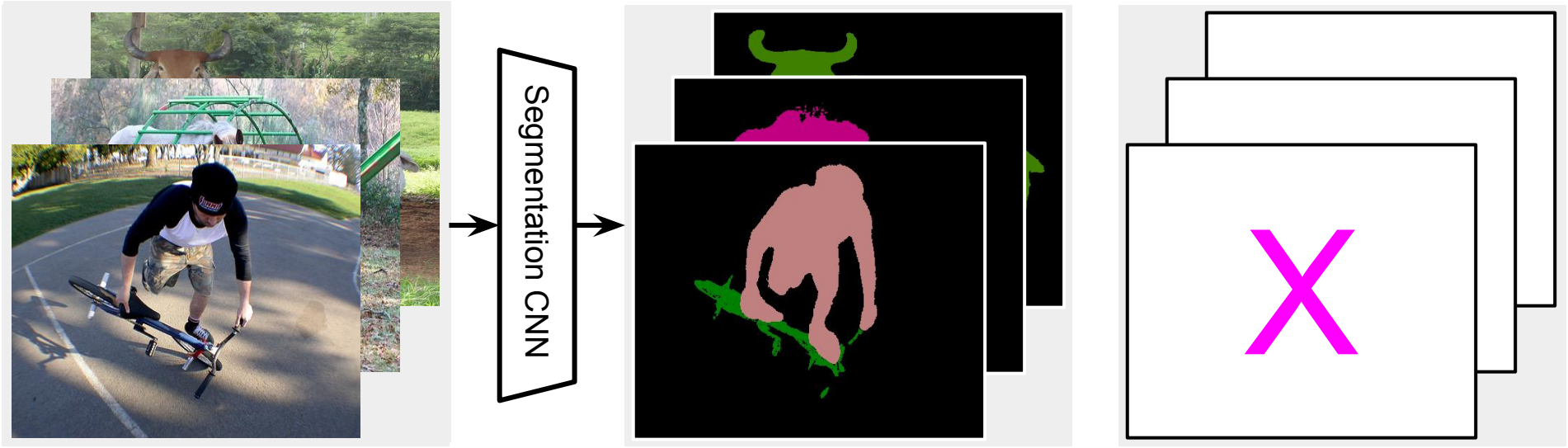
Stella X. Yu

Berkeley
UNIVERSITY OF CALIFORNIA

Task of Semantic Segmentation:
Put Pixels into Semantic Categories



Task of Unsupervised Semantic Segmentation:
Put Pixels into Groups without Any Labeled Supervision



Two Approaches to Predict Pixel Labels from Groupings

Images

Segmentation CNN

Pixel Features

1. Clustering

Group 3

Group 2

Group 1

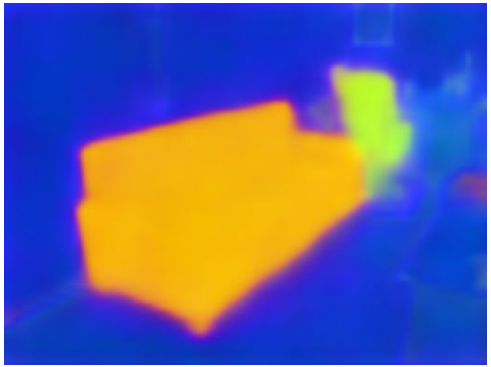
2. Nearest Neighbor Search

1. *Invariant Information Clustering for Unsupervised Image Classification and Segmentation. Ji et al. ICCV 2019.*
2. *SegSort: Segmentation by Discriminative Sorting of Segments*. Hwang et al. ICCV 2019.

Our Model by Feature Learning: Predict Labels from Retrieved Segments

Images

Segmentation CNN



Pixel Features

Query

Retrievals

Sofa

Chair

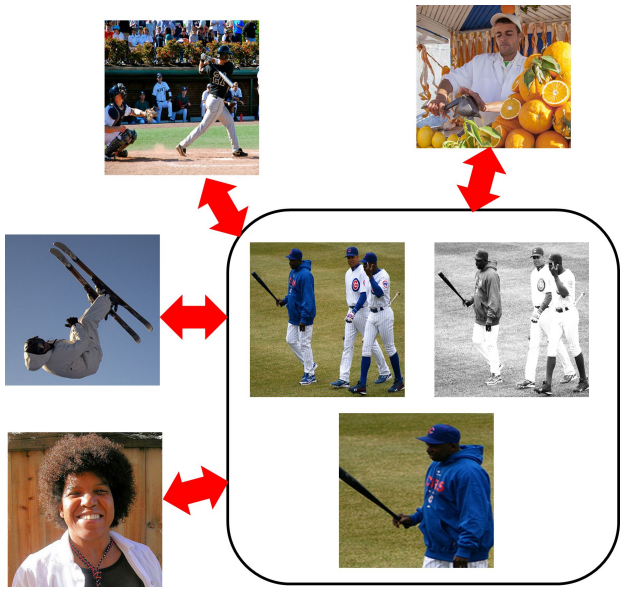
Sofa

Sofa

Chair

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

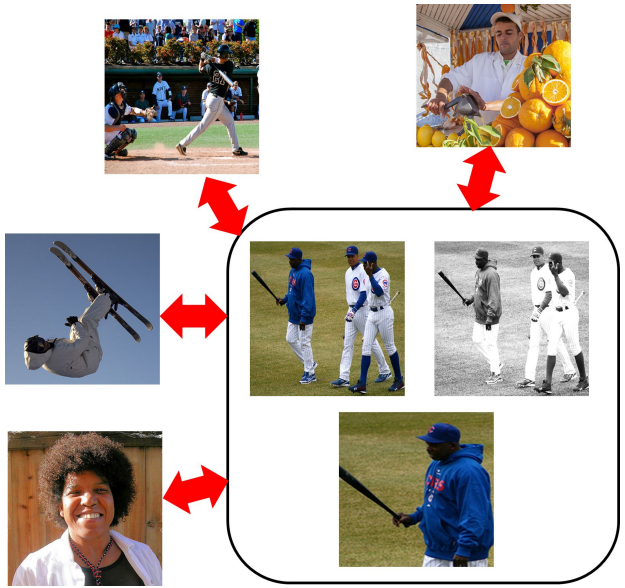
Contrast **images** disregarding visual change



Revisiting Contrastive Methods for Unsupervised Learning of Visual Representations. Gansbeke et al. NeuRIPS 2021.

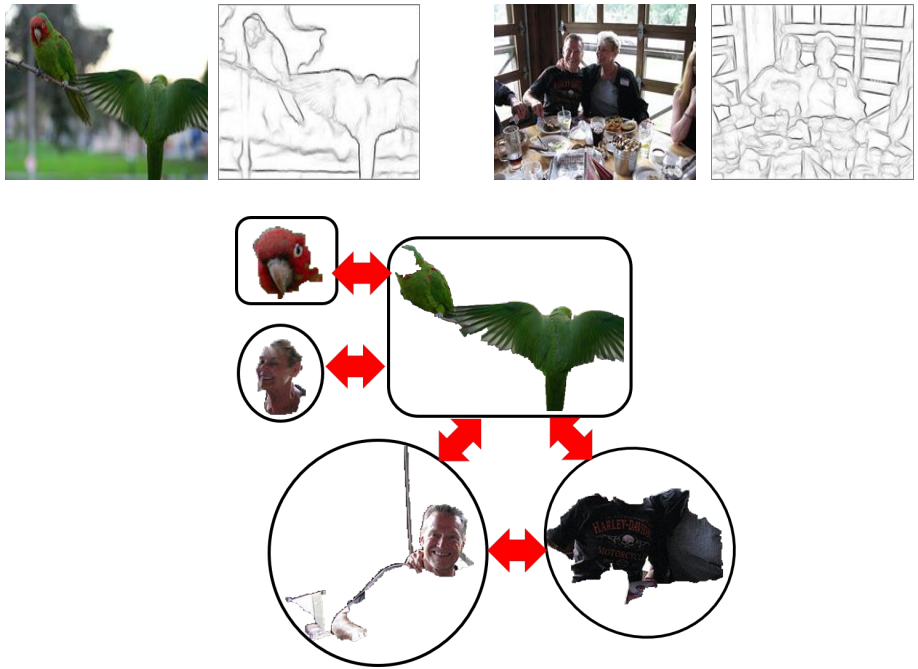
Current Feature Learning Methods: Contrast Image-Image vs. Pixel-Segment

Contrast **images** disregarding visual change



Revisiting Contrastive Methods for Unsupervised Learning of Visual Representations. Gansbeke et al. NeuRIPS 2021.

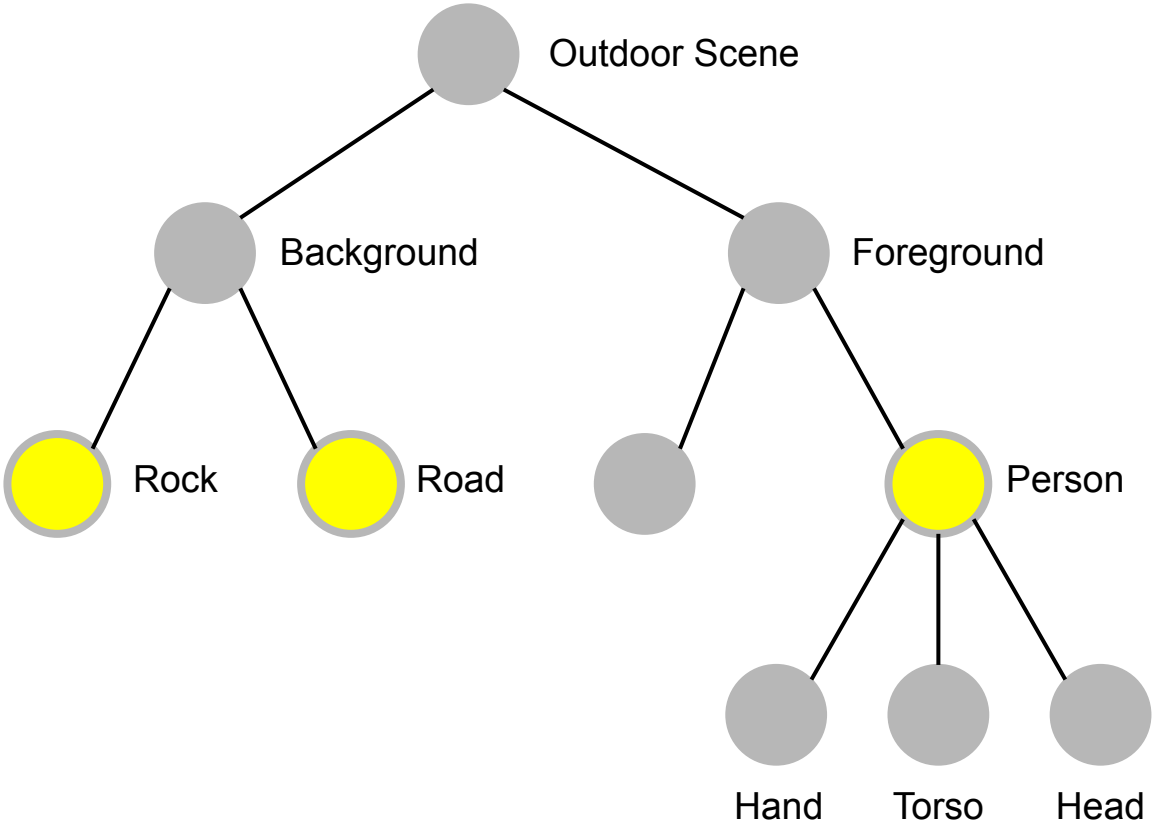
Contrast pixels with **regions** w.r.t low-level visual cues



SegSort: Segmentation by Discriminative Sorting of Segments. Hwang et al. ICCV 2019.

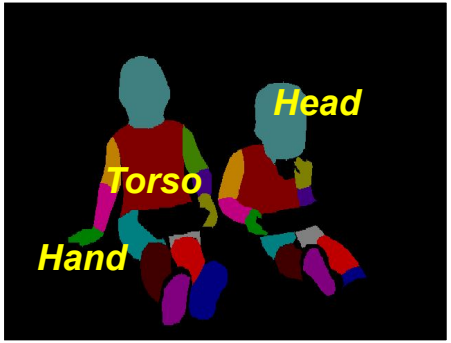
Semantics Intrinsically Has Multiple Levels of Granularity

Coarse-grained Categories

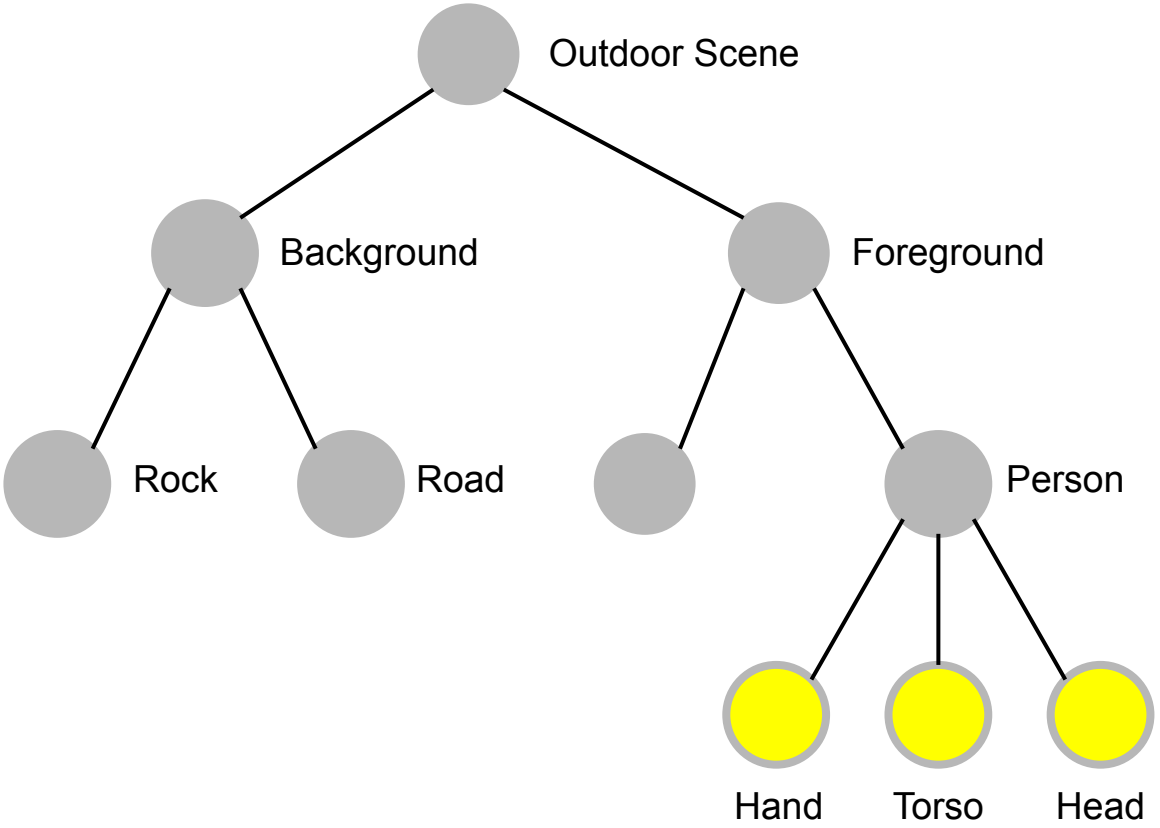


Microsoft coco: Common objects in context. Lin et al. ECCV 2014.

Semantics Intrinsically Has Multiple Levels of Granularity



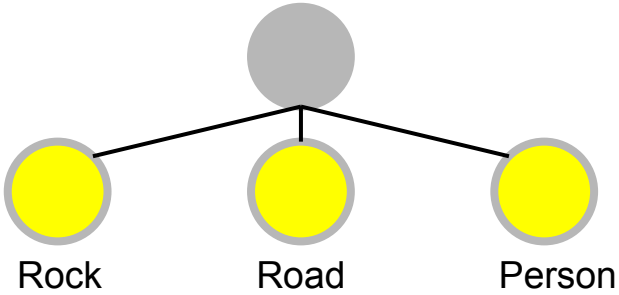
Fine-grained Categories



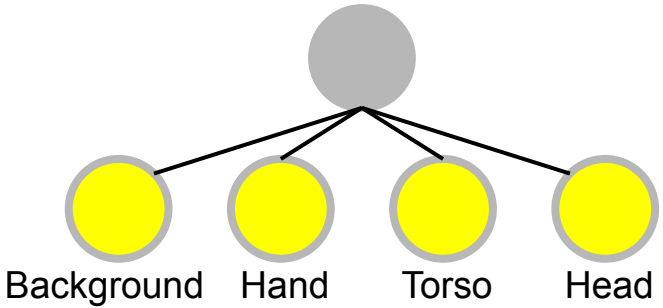
Densepose: Dense human pose estimation in the wild. G¨uler et al. CVPR 2018.

Semantics Intrinsically Has Multiple Levels of Granularity

Most existing methods:
avoid ambiguity / presume a granularity

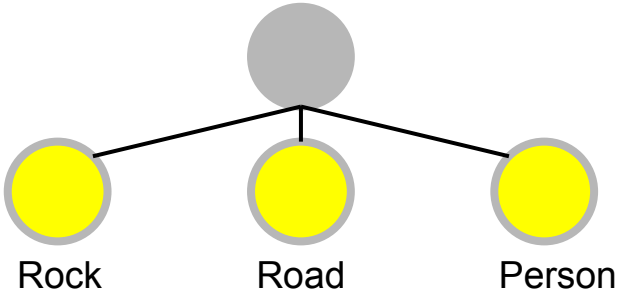


or

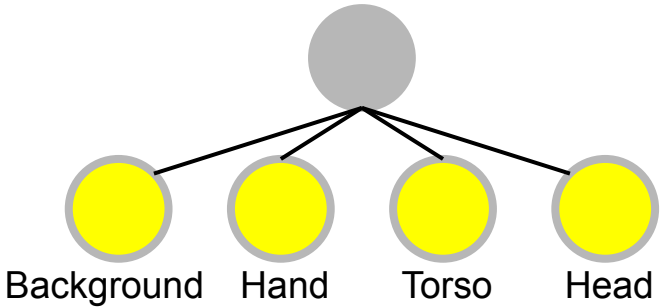


Semantics Intrinsically Has Multiple Levels of Granularity

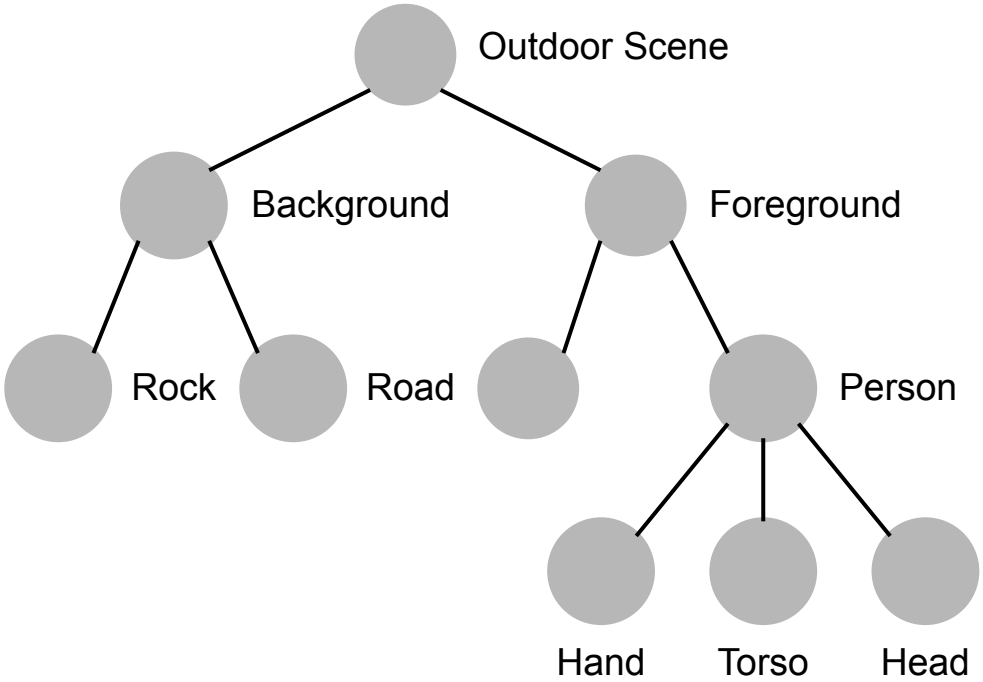
Most existing methods:
avoid ambiguity / presume a granularity



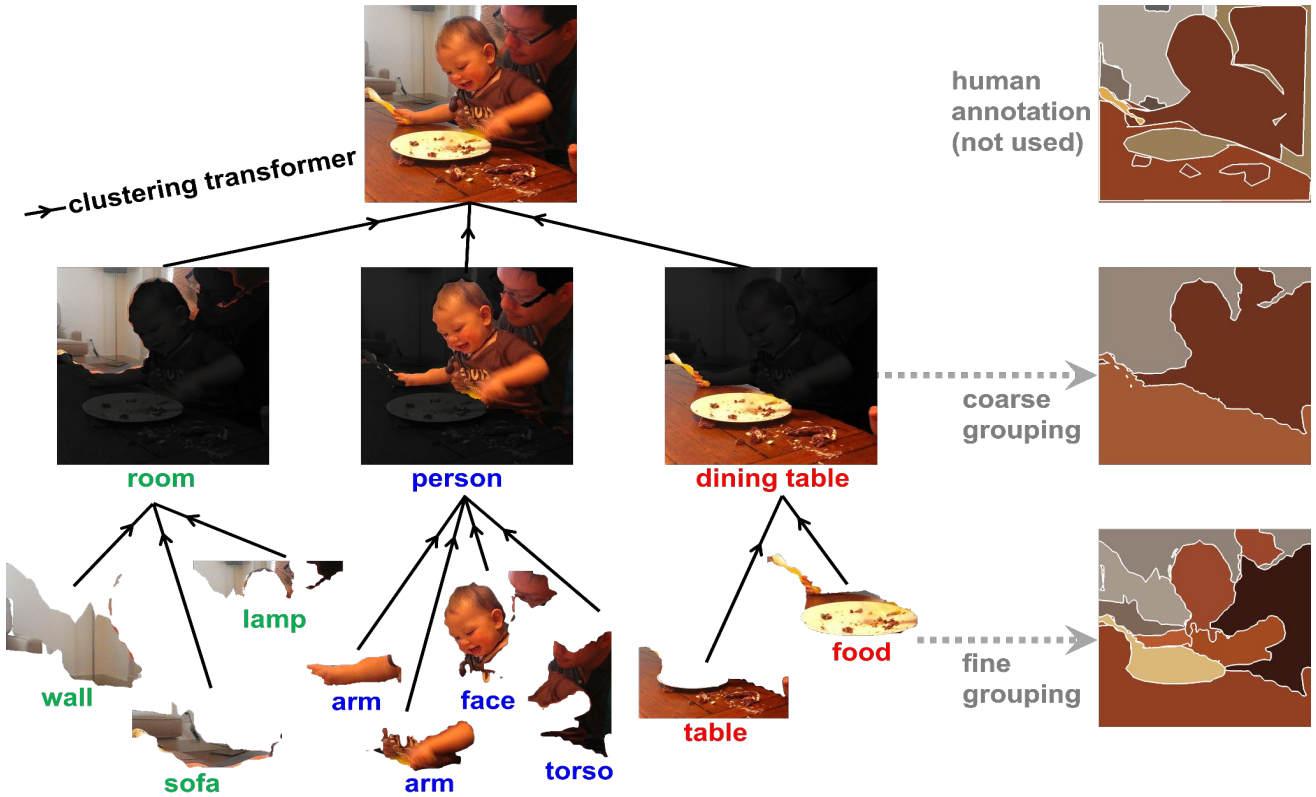
or



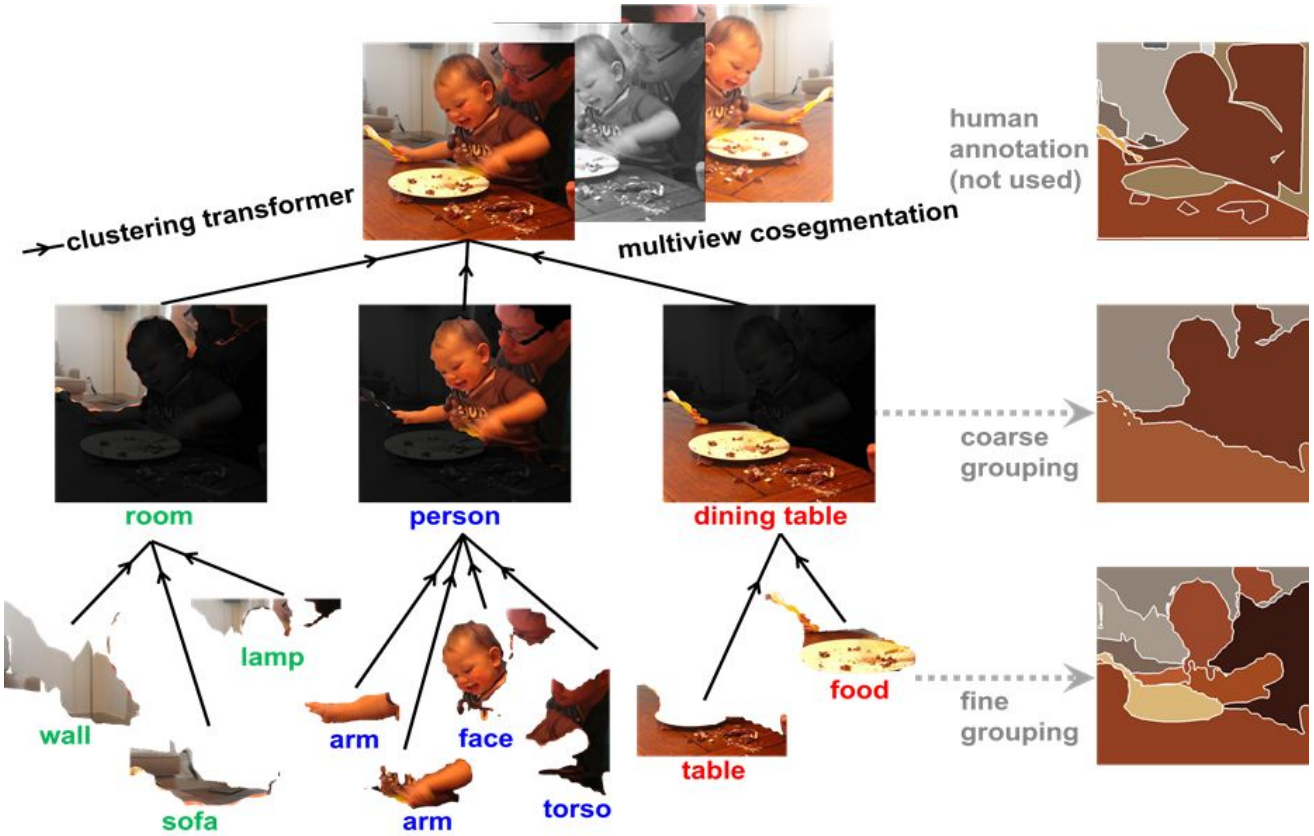
Our idea: embrace multiple levels of granularity



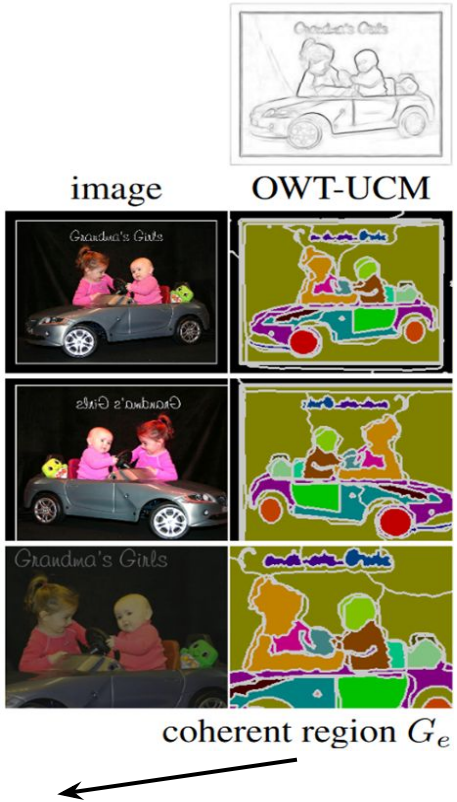
Face and Body are Parts of a Whole in the Visual Scene



Babies Appear Different but Have the Same Semantics

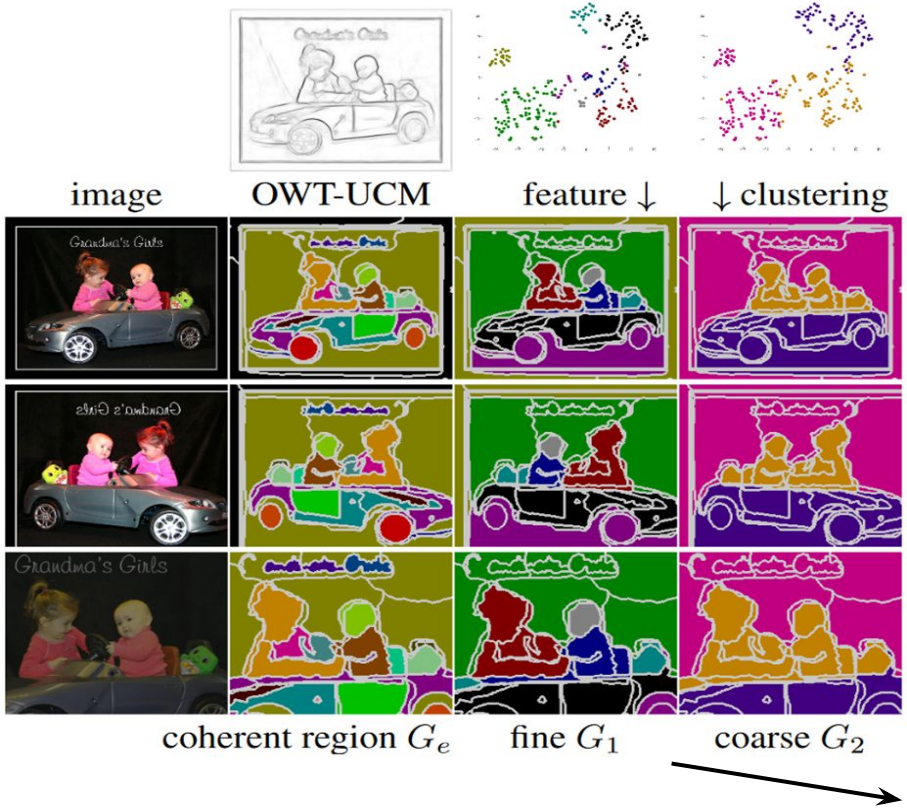


Invariance: Multiview Cosegmentation



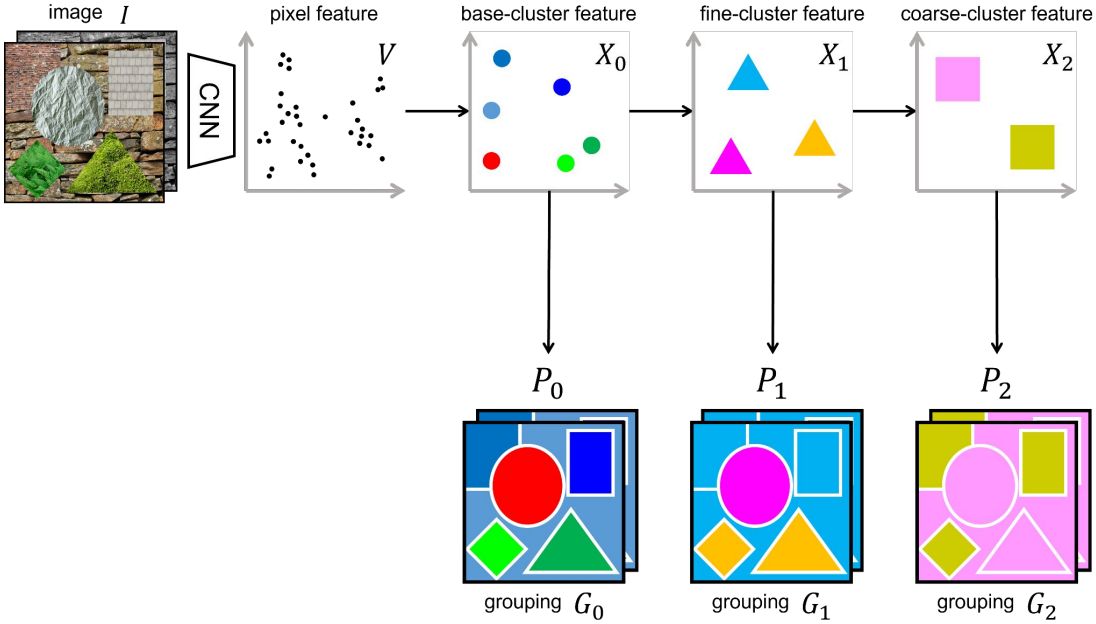
Ground features by visual appearance and correspondence

Invariance: Multiview Cosegmentation



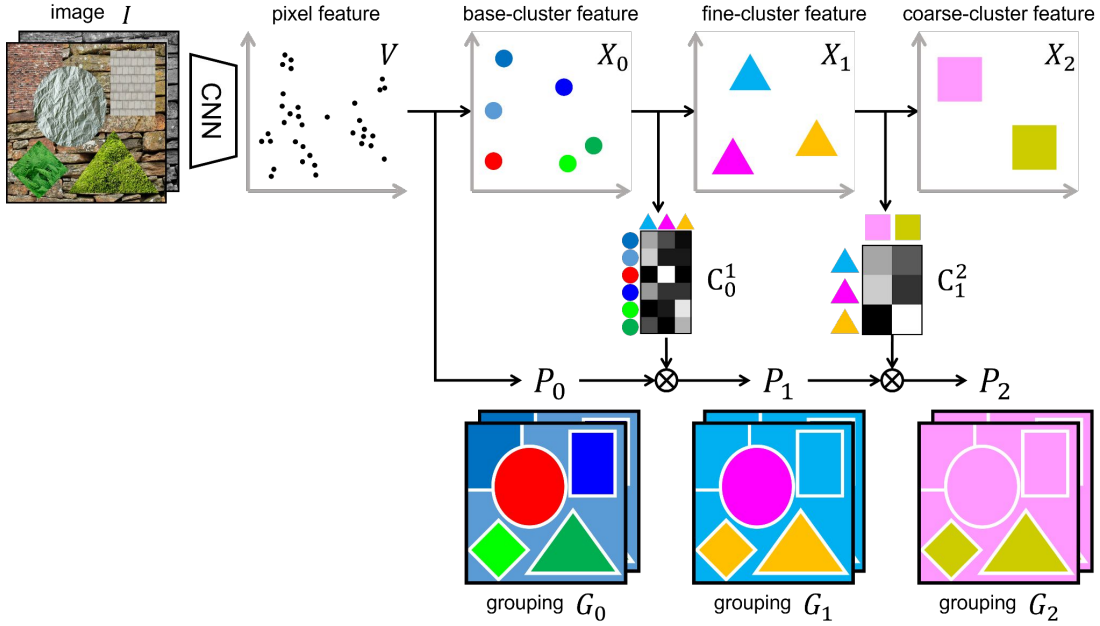
Regularize features by multi-scale grouping

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

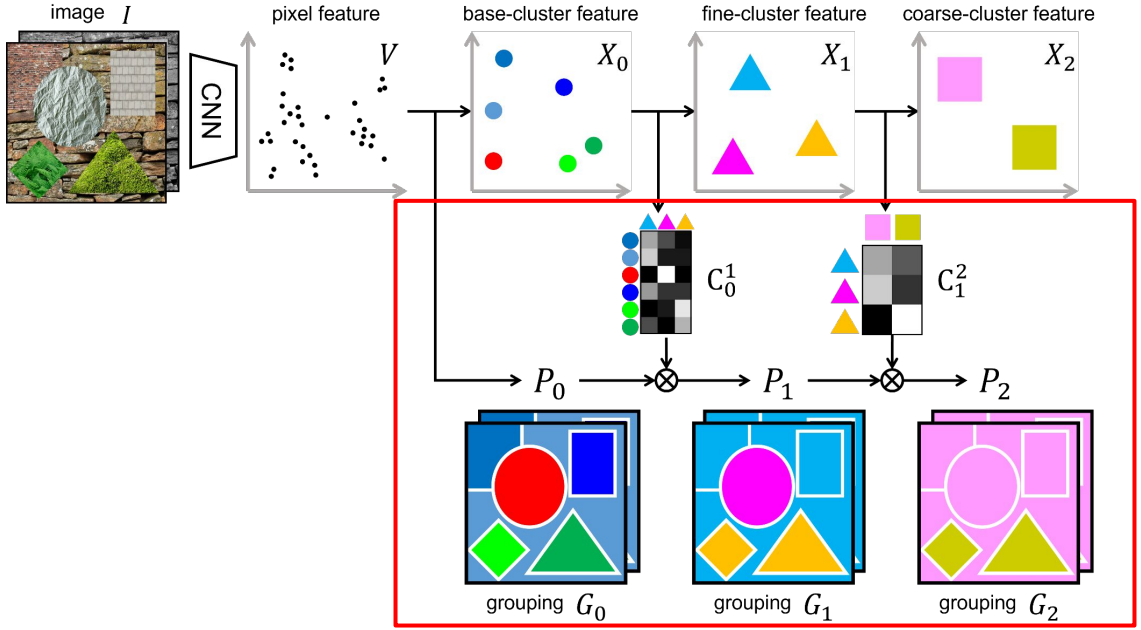


Grouping Probability at Level l: $P_l(a) = \text{Prob}(G_l = a|x)$

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

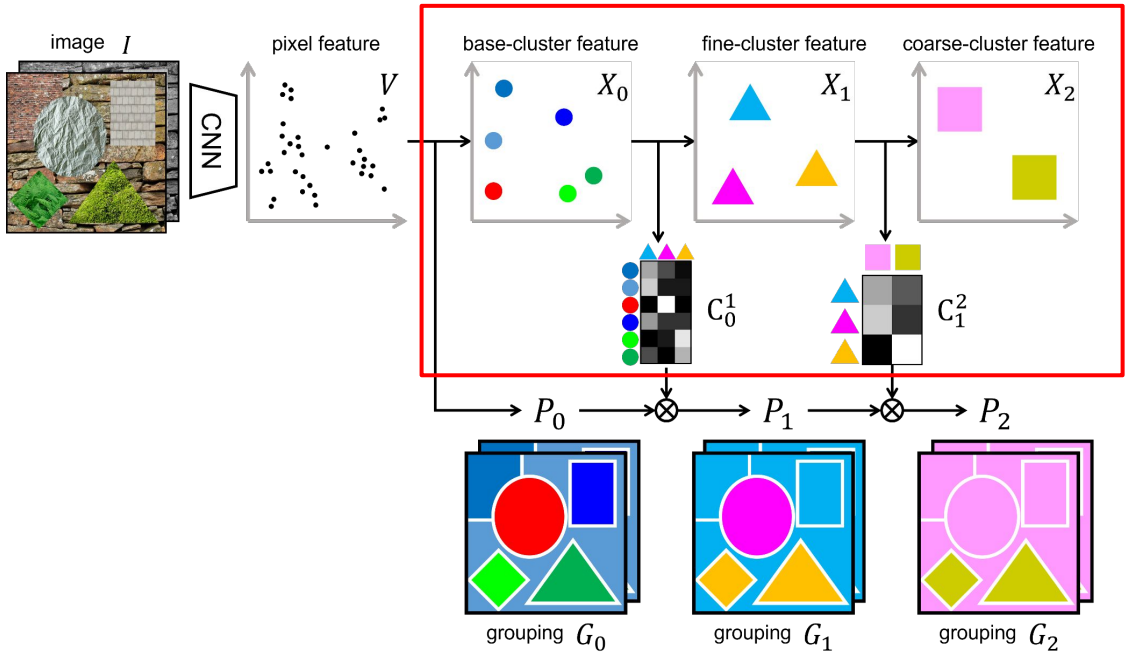


Grouping Probability at Level l: $P_l(a) = \text{Prob}(G_l = a|x)$

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

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

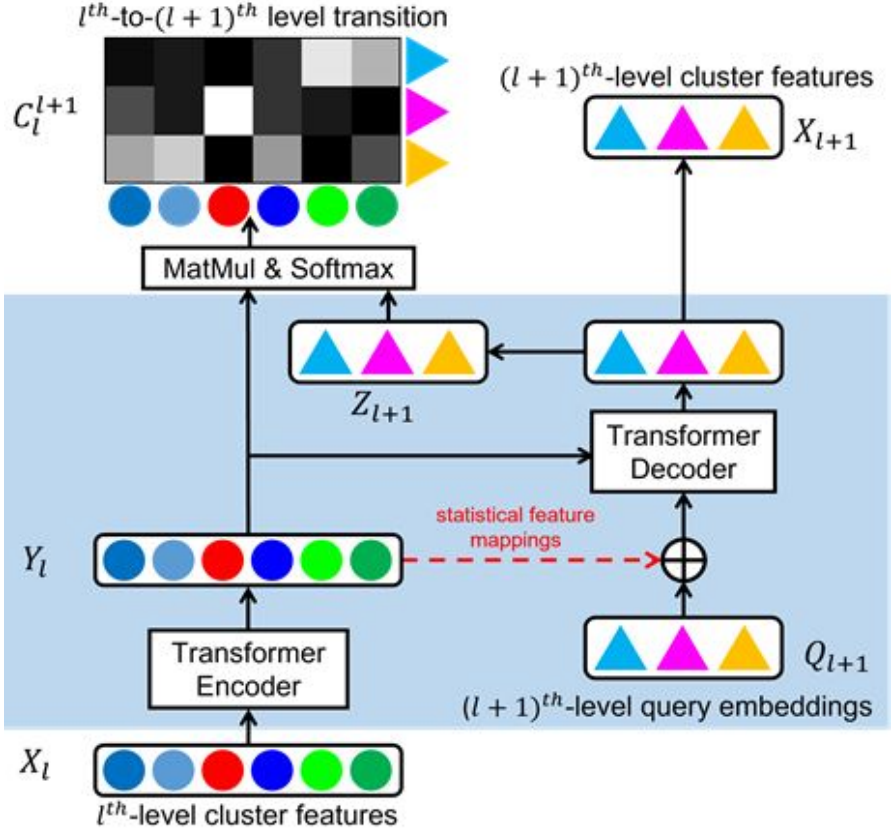


Grouping Probability at Level l: $P_l(a) = \text{Prob}(G_l = a|x)$

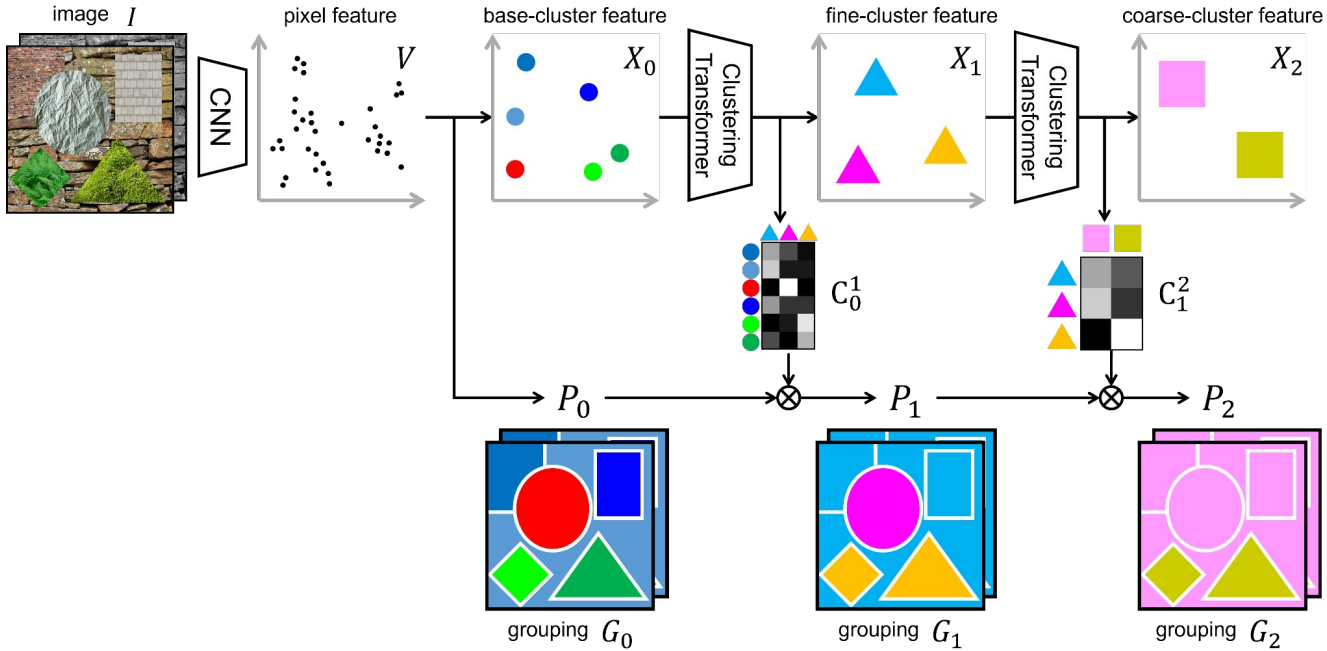
Transition Probability to Level $l+1$: $C_l^{l+1}(a,b) = \text{Prob}(G_{l+1} = b|G_l = a)$

Grouping Assignment at Level $l+1$: $P_{l+1} = P_l \times C_l^{l+1} = P_0 \times C_0^1 \times \cdots \times C_l^{l+1}$

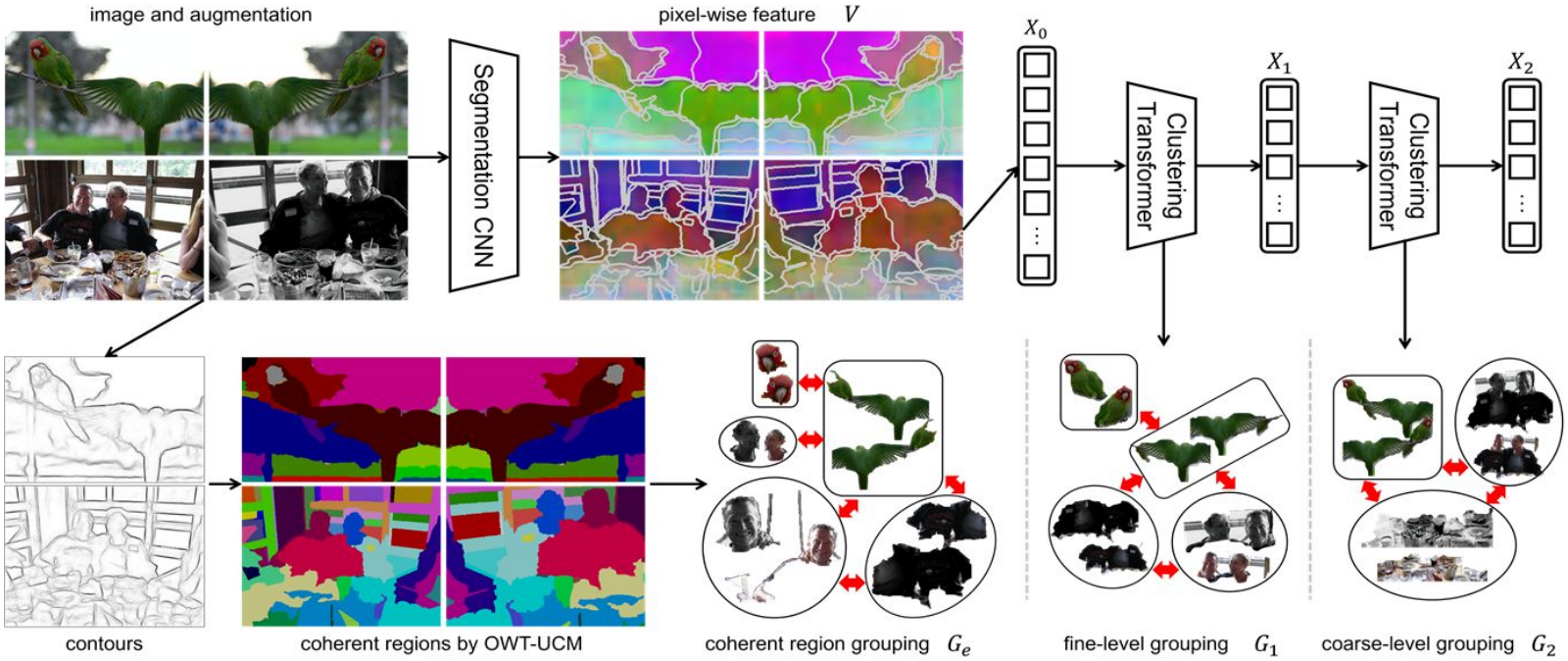
Multi-scale Consistency: Clustering Transformer



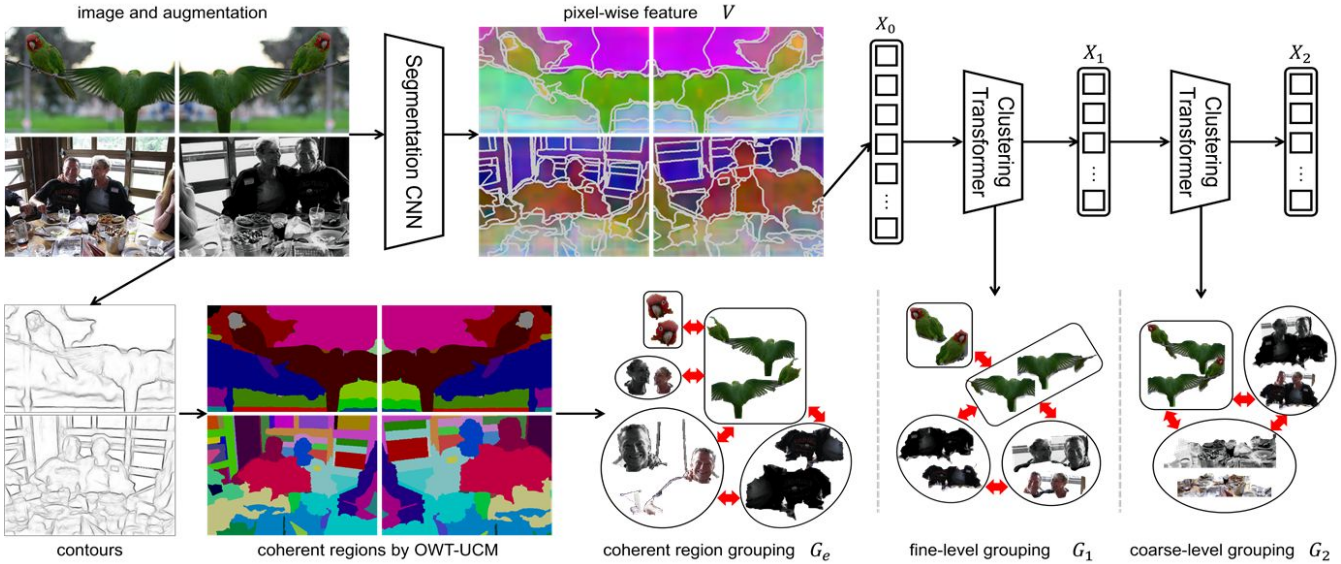
Multi-scale Consistency: Clustering Transformer



Our Hierarchical Segment Grouping Model



Our Hierarchical Segment Grouping Model

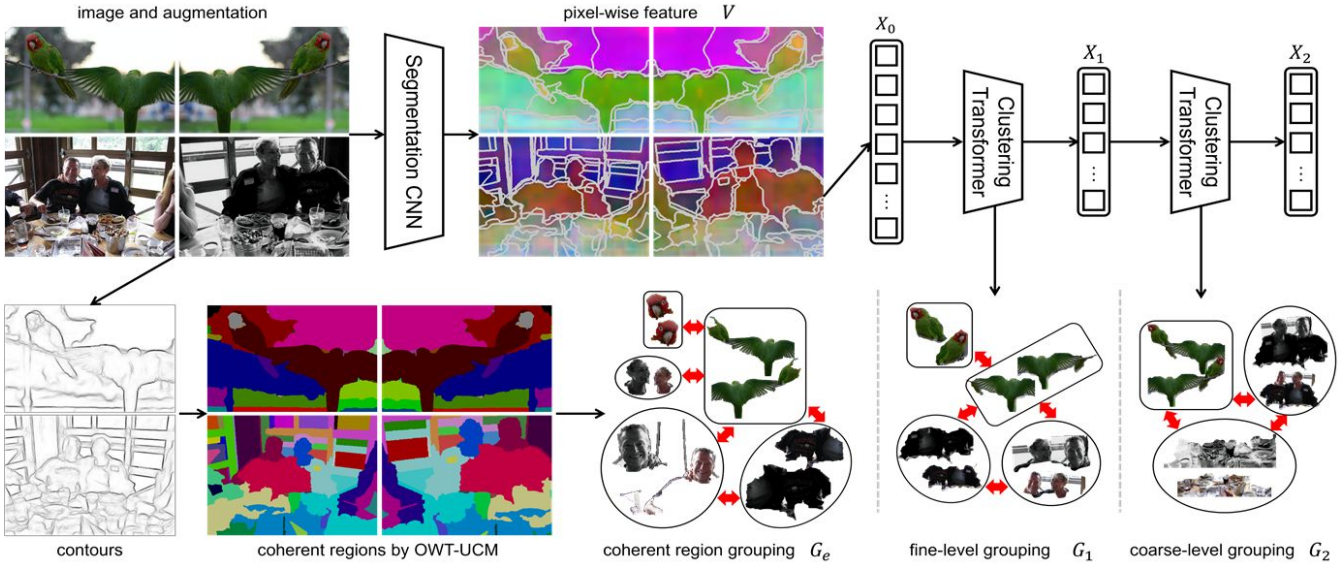


$$L(f) = \lambda_E L_f(G_e) + \lambda_F \sum_{l \geq 1} L_f(G_l) + \lambda_G L_g$$

Pixel-segment contrast loss:

1. Ground features by visual appearance
2. Enforce correspondence across views

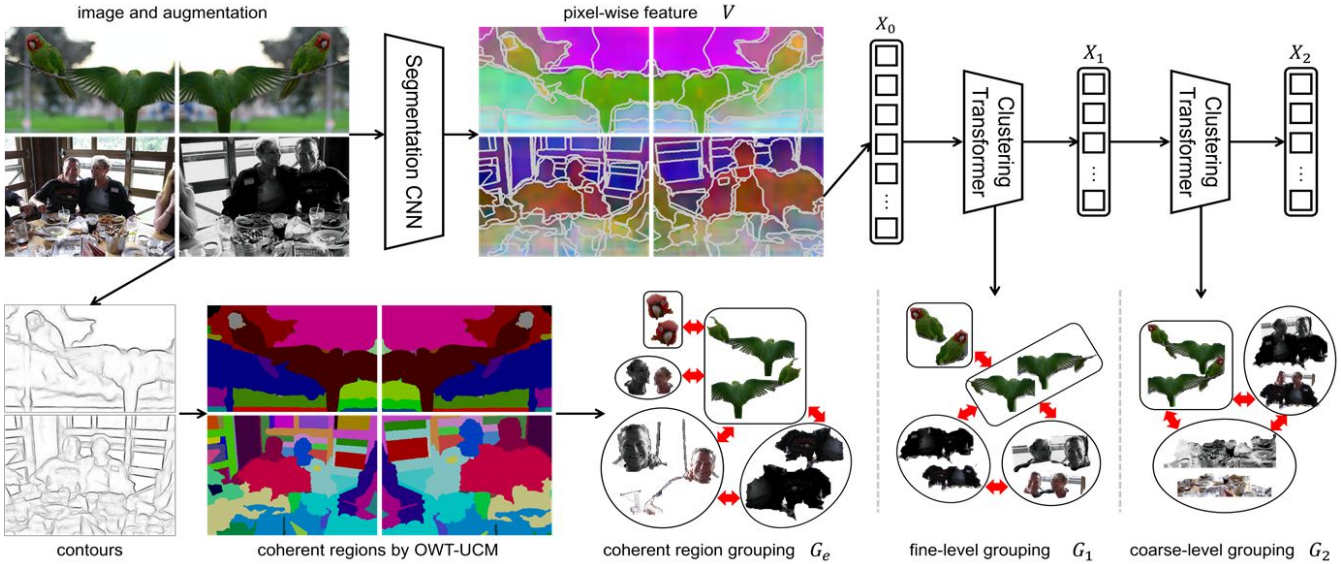
Our Hierarchical Segment Grouping Model



$$L(f) = \lambda_E L_f(G_e) + \lambda_F \sum_{l \geq 1} L_f(G_l) + \lambda_G L_g$$

Pixel-segment contrast loss:
Regularize features by consistent hierarchy

Our Hierarchical Segment Grouping Model

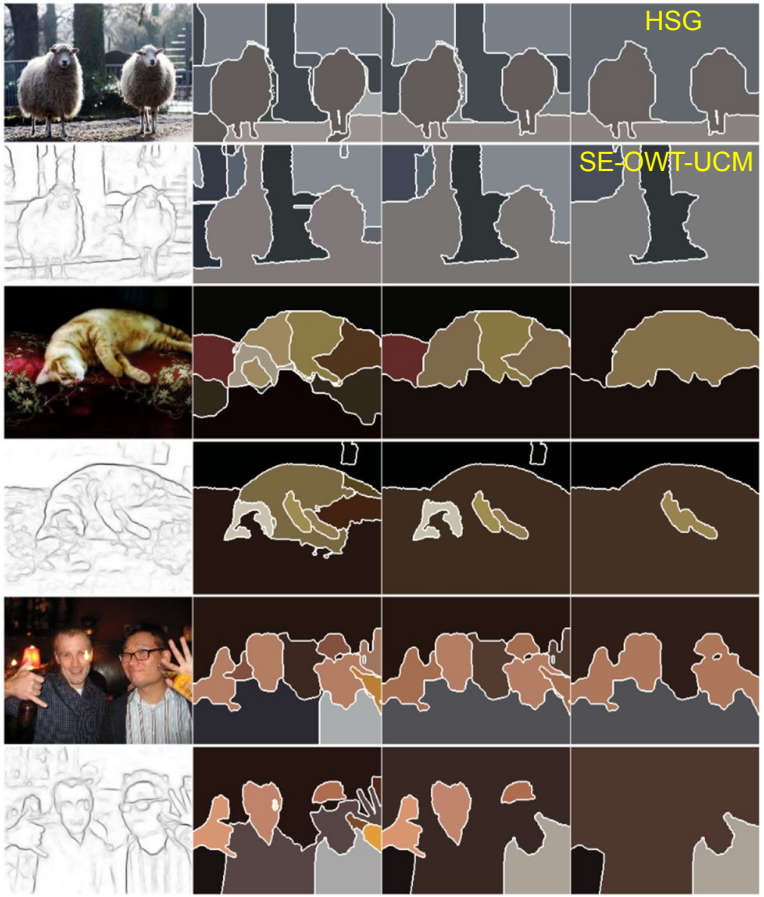


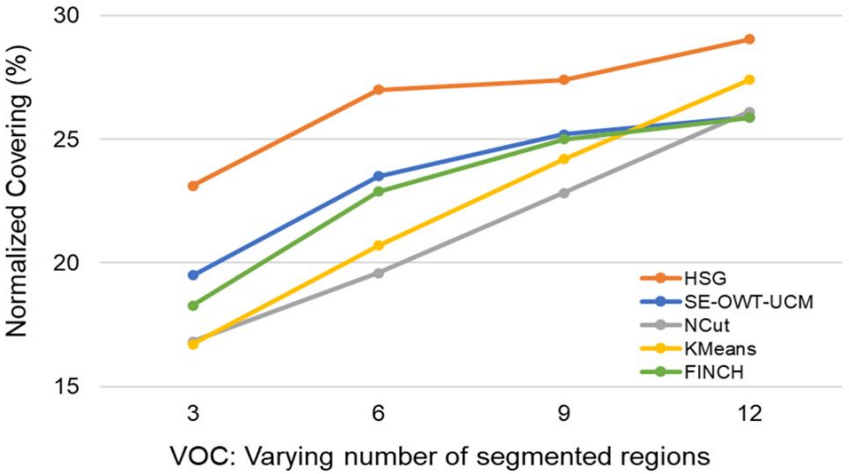
$$L(f) = \lambda_E L_f(G_e) + \lambda_F \sum_{l \geq 1} L_f(G_l) + \lambda_G L_g$$

Goodness of Grouping:
Desire balanced, compact, distinctive clusters

Graph Clustering with Graph Neural Networks. Tsitsulin et al. ArXiv 2020.

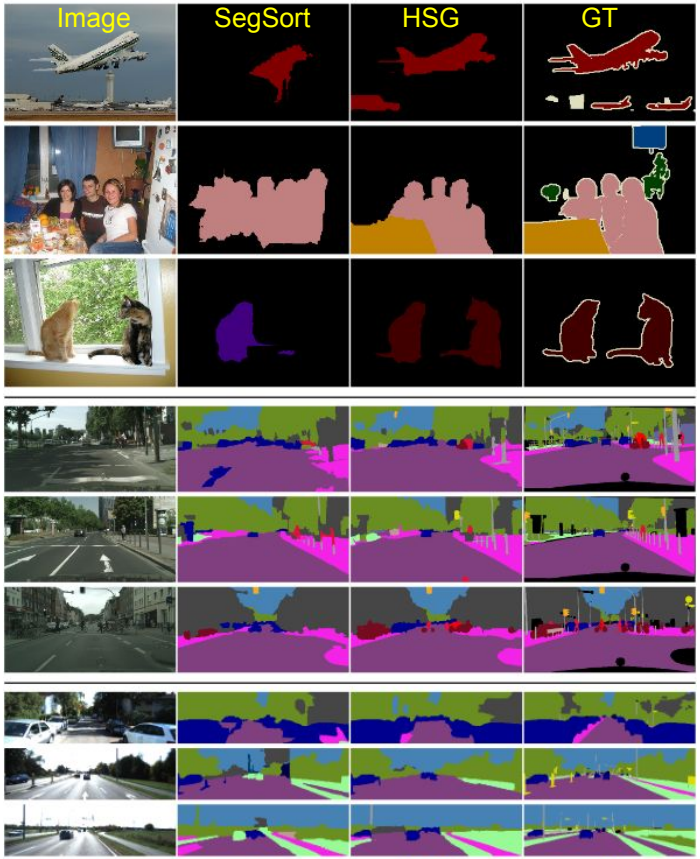
1. First Unsupervised Hierarchical Semantic Segmentation





$$\mathrm{NFCovering}(S' \to S_{fg}) = \frac{1}{|S_{fg}|} \sum_{R \in S_{fg}} \max_{R' \in S'} \frac{|R \cap R'|}{|R \cup R'|}$$

2. SOTA on Unsupervised Semantic Segmentation



Training set	MSCOCO		Cityscapes		KITTI-STEP	
Validation set	VOC		Cityscapes		KITTI-STEP	
Method	mIoU	Acc.	mIoU	Acc.	mIoU	Acc.
Moco [20]	28.1	-	15.3	69.5	13.7	60.3
DenseCL [60]	35.1	-	12.7	64.2	9.3	47.6
Revisit [56]	35.1	-	17.1	71.7	17.0	65.0
SegSort [26]	11.7	75.1	24.6	81.9	19.2	69.8
Our HSG	**41.9**	**85.7**	**32.5**	**86.0**	**21.7**	**73.8**

3. Unsupervised Visual Context Retrievals across Granularity Levels

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