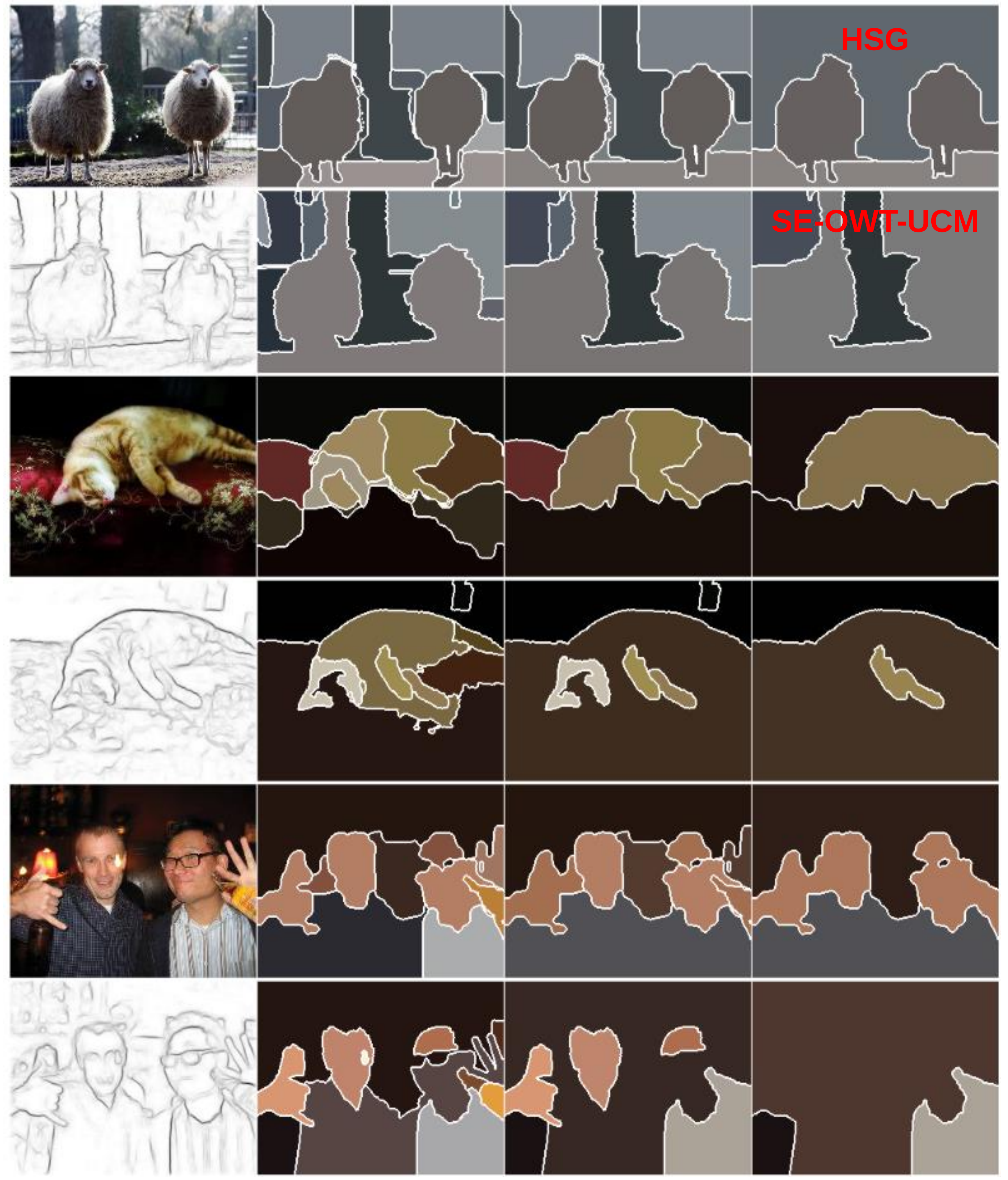
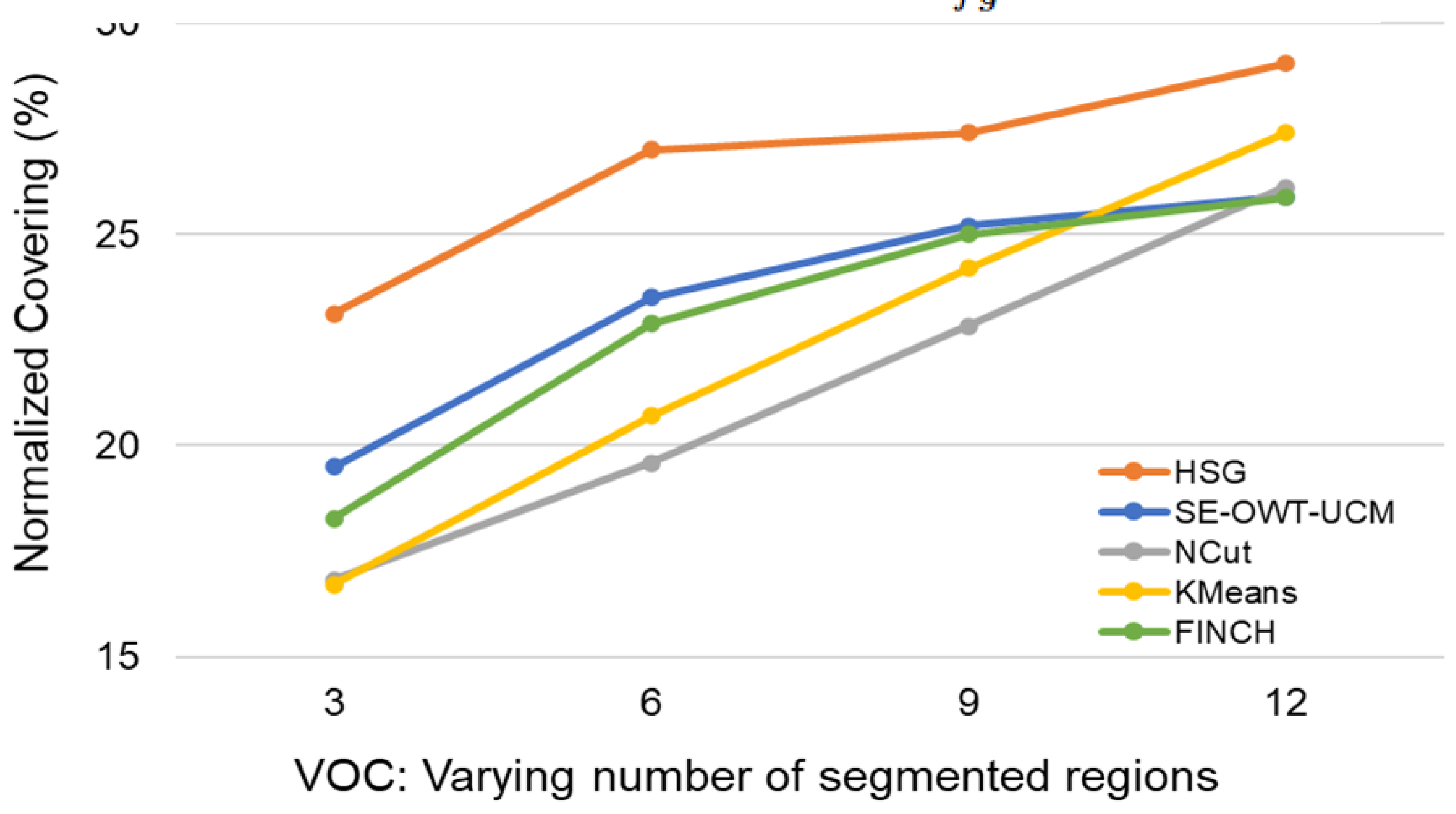
Unsupervised Hierarchical Semantic Segmentation w/ Multiview Cosegmentation & Clustering Transformers

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

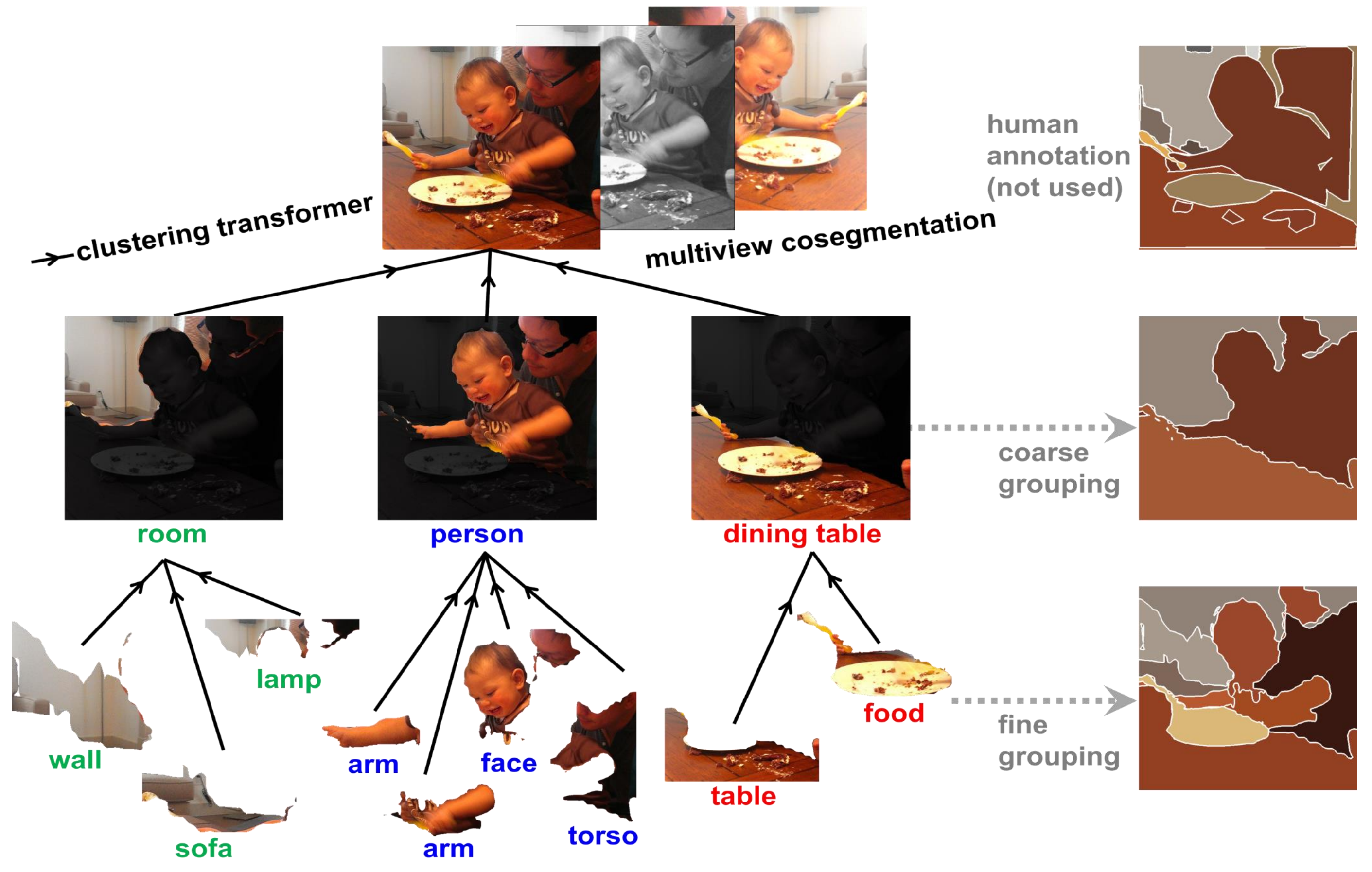
Unsupervised Hierarchical Segmentation



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



Contributions



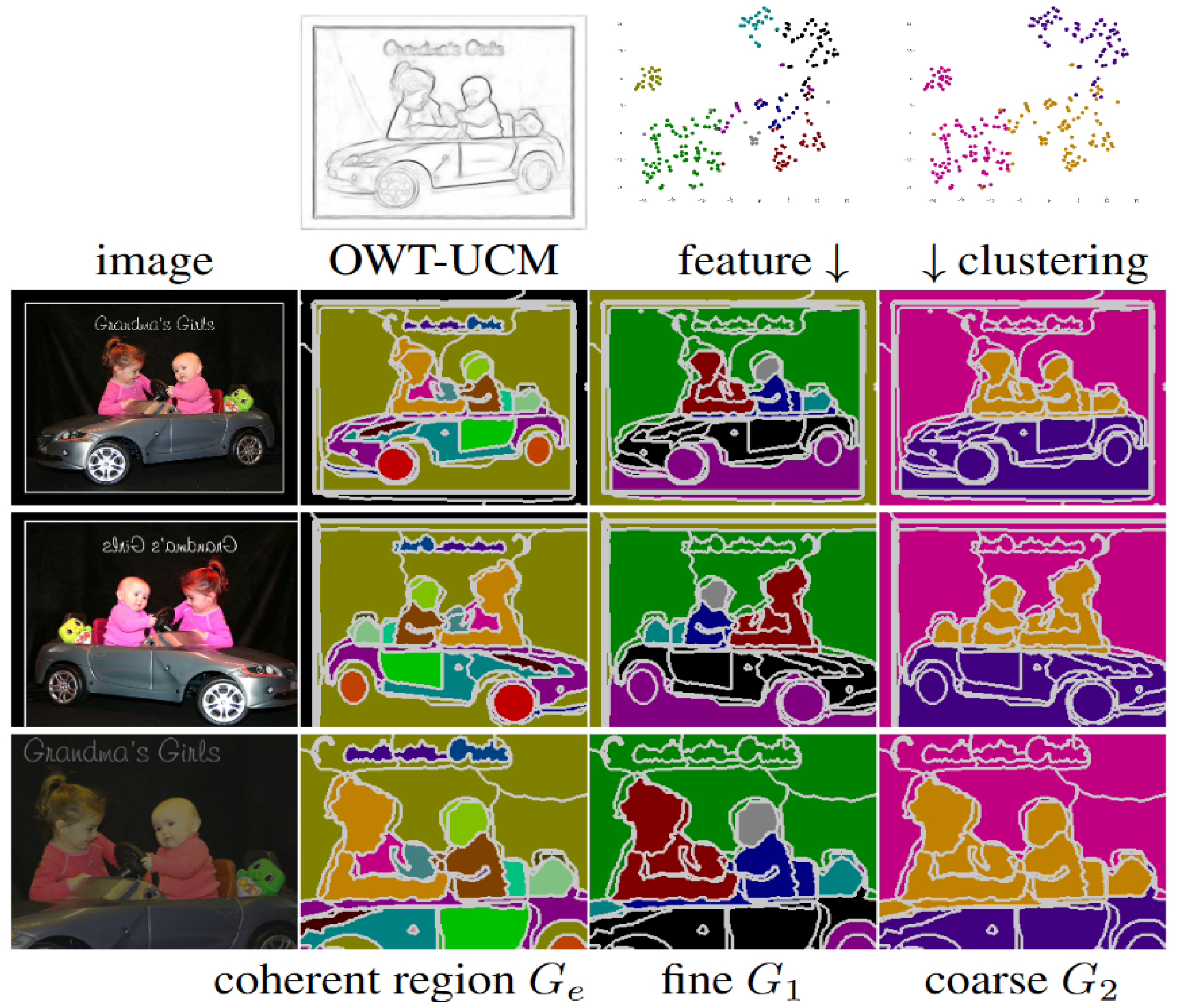
First unsupervised hierarchical semantic segmentation

First feature learning that embraces scale ambiguity

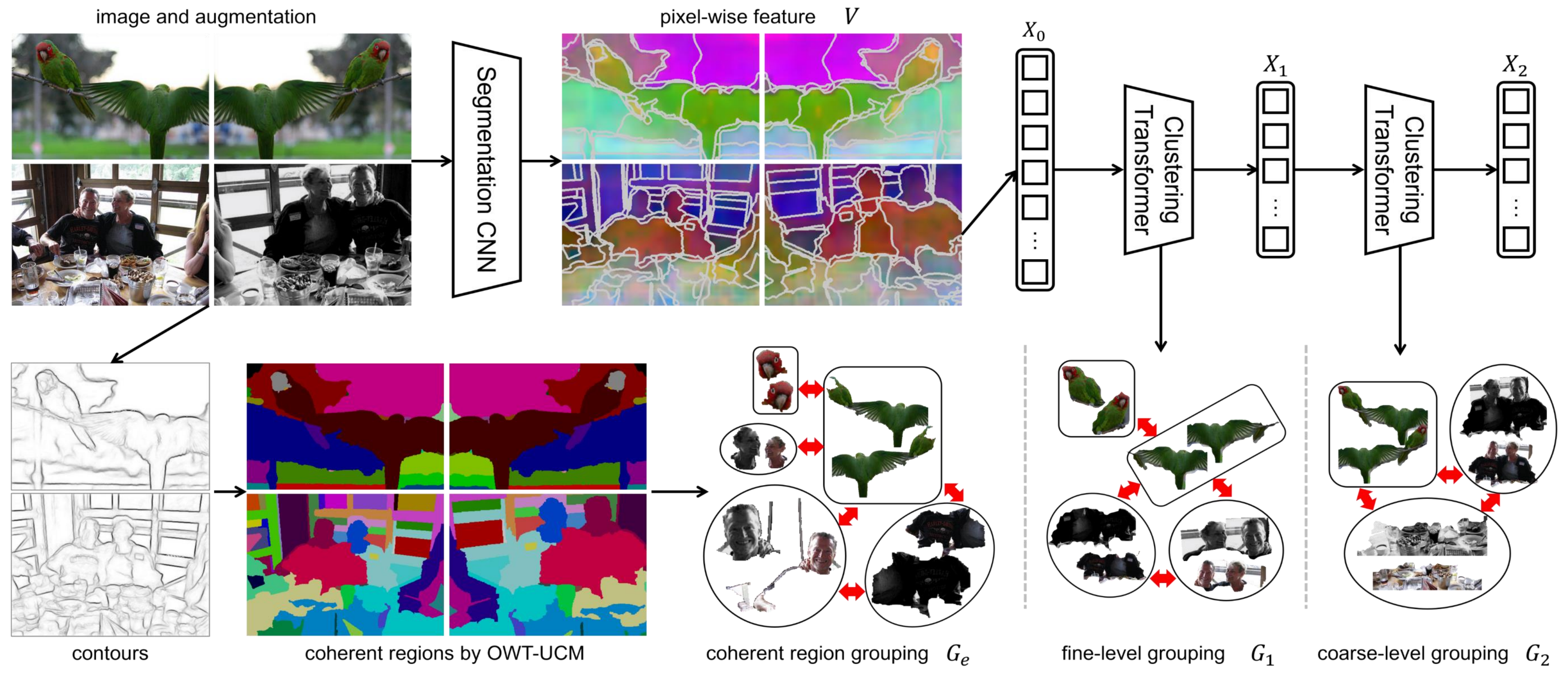
SOTA on unsupervised semantic segmentation

Invariance: Multiview Cosegmentation

Babies appear different but have the same semantics



Hierarchical Segment Grouping Model



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

Contrastive feature loss $L_f(G_e)$ grounds features by visual appearance

Contrastive feature loss $L_f(G_l)$ regularizes features by consistent hierarchy

Goodness of grouping L_g desires balanced, compact, distinctive clusters

Consistency: Clustering Transformers

Face and body are parts of a whole in the visual scene

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

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

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

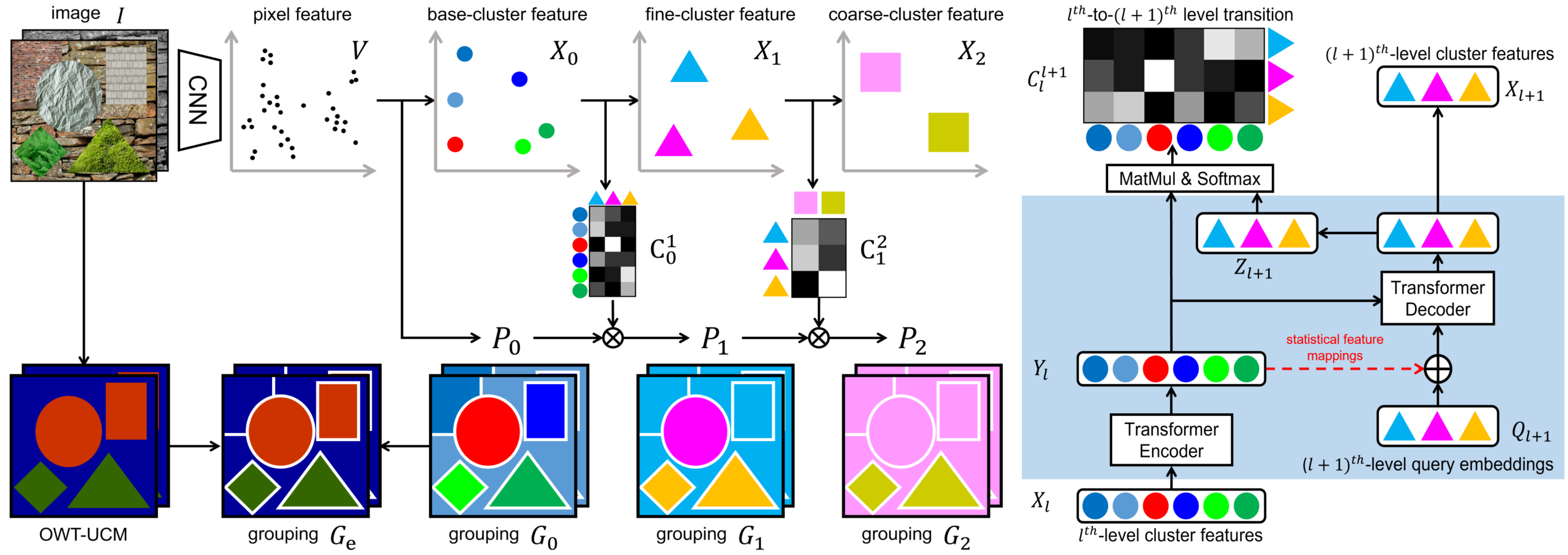
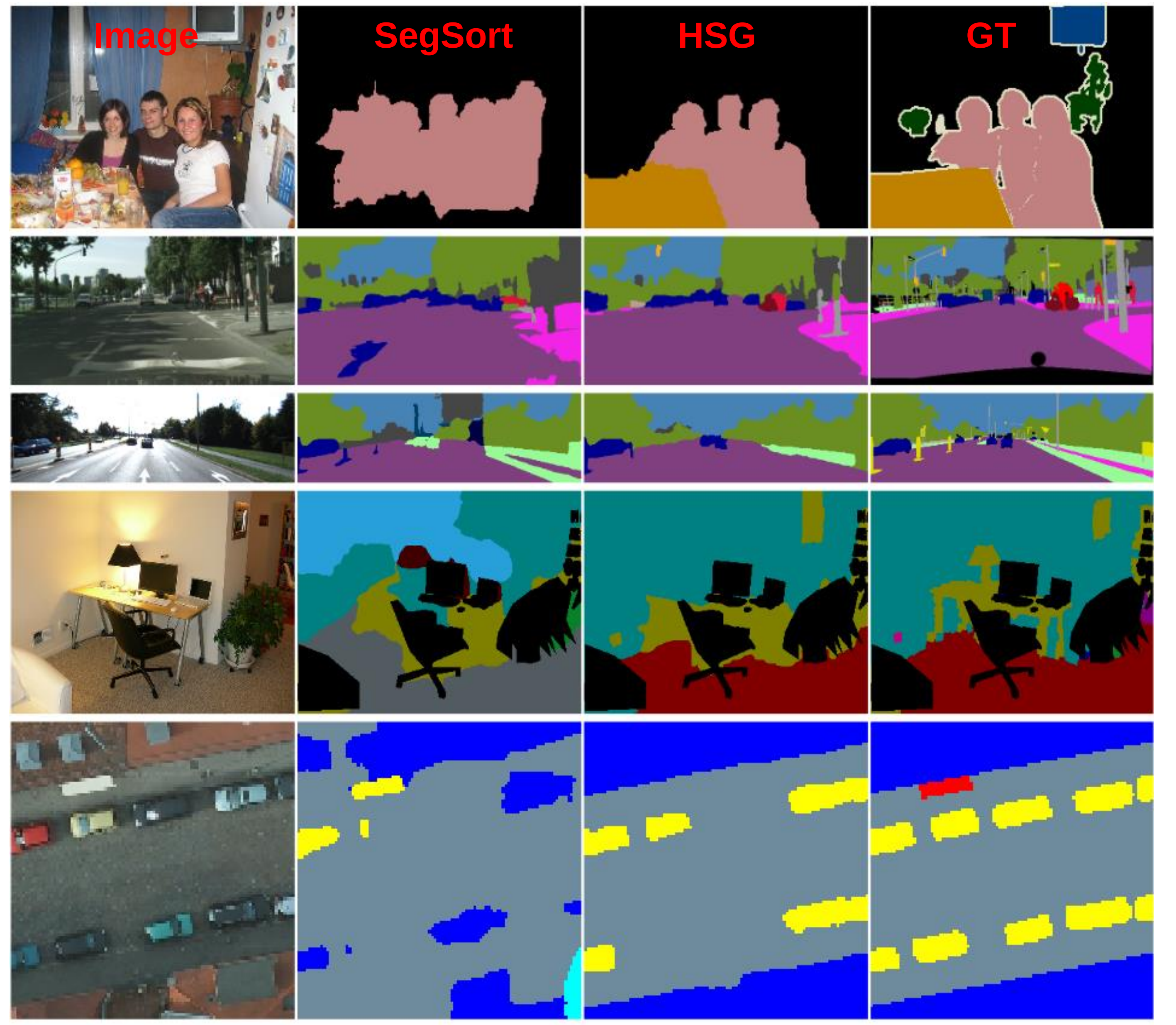


image I
CNN
pixel feature V
base-cluster feature X_0
fine-cluster feature X_1
coarse-cluster feature X_2
C_0^1
C_1^2
P_0
P_1
P_2
OWT-UCM
grouping G_e
grouping G_0
grouping G_1
grouping G_2
l^{th}-to-$(l+1)^{th}$ level transition
C_l^{l+1}
$(l+1)^{th}$-level cluster features
X_{l+1}
MatMul & Softmax
Z_{l+1}
Transformer Decoder
statistical feature mappings
Y_l
Transformer Encoder
Q_{l+1}
$(l+1)^{th}$-level query embeddings
X_l
l^{th}-level cluster features

Unsupervised Semantic Segmentation



Train	COCO	Cityscapes	KITTI
Test	VOC	Cityscapes	KITTI
Moco	28.1	15.3	13.7
DenseCL	35.1	12.7	9.3
Revisit	35.1	17.1	17.0
SegSort	11.7	24.6	19.2
Our HSG	**41.9**	**32.5**	**21.7**

Train & Test	COCO-stuff	Potsdam
DeepCluster	19.9	29.2
IIC	27.7	45.4
AC	30.8	49.3
SegSort	49.9	59.0
Our HSG	**57.6**	**67.4**

Unsupervised Contextual Retrievals