Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers

Tsung-Wei Ke Jyh-Jing Hwang Yunhui Guo Xudong Wang Stella X. Yu
UC Berkeley / ICSI

Abstract

Unsupervised semantic segmentation aims to discover groupings within and across images that capture object- and view-invariance of a category without external supervision. Grouping naturally has levels of granularity, creating ambiguity in unsupervised segmentation. Existing methods avoid this ambiguity and treat it as a factor outside modeling, whereas we embrace it and desire hierarchical grouping consistency for unsupervised segmentation.

We approach unsupervised segmentation as a pixel-wise feature learning problem. Our idea is that a good representation shall reveal not just a particular level of grouping, but any level of grouping in a consistent and predictable manner. We enforce spatial consistency of grouping and bootstrap feature learning with co-segmentation among multiple views of the same image, and enforce semantic consistency across the grouping hierarchy with clustering transformers between coarse- and fine-grained features.

We deliver the first data-driven unsupervised hierarchical semantic segmentation method called Hierarchical Segment Grouping (HSG). Capturing visual similarity and statistical co-occurrences, HSG also outperforms existing unsupervised segmentation methods by a large margin on five major object- and scene-centric benchmarks.

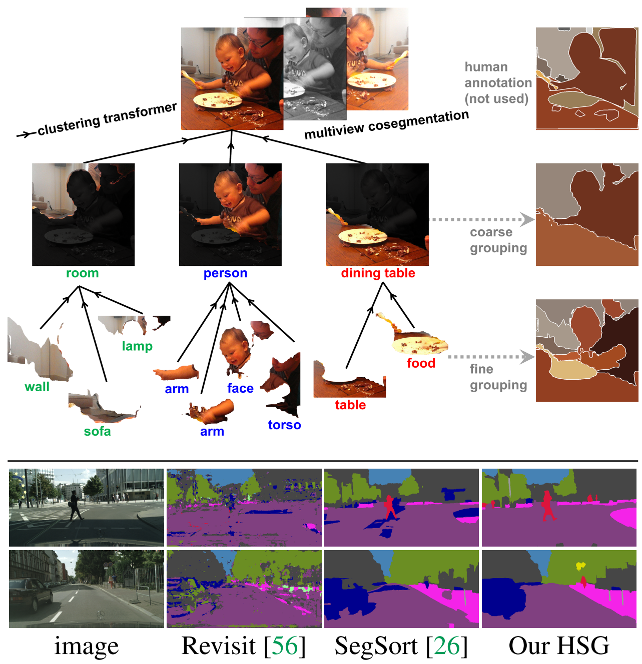

Figure 1. We develop an unsupervised semantic segmentation method by embracing the ambiguity of grouping granularity and desiring hierarchical grouping consistency for unsupervised segmentation. **Top:** We formulate it as a pixel-wise feature learning problem, such that a good feature must be able to best reveal any level of grouping in a consistent and predictable manner. We bootstrap feature learning from multiview cosegmentation and enforce grouping consistency with clustering transformers. **Bottom:** Our method can not only deliver *hierarchical* semantic segmentation, but also outperform the state-of-the-art unsupervised segmentation methods by a large margin. Shown are sample Cityscapes results.

1. Introduction

Semantic segmentation requires figuring out the semantic category for each pixel in an image. Learning such a segmenter from unlabeled data is particularly challenging, as neither pixel groupings nor semantic categories are known.

If pixel groupings are known, semantic segmentation is reduced to an unsupervised image (segment) recognition problem, to which contrast learning methods [9, 20, 59, 62] could apply, on computed segments instead of images.

If semantic categories are known, semantic segmentation is reduced to a weakly supervised segmentation problem with coarse annotations of image-level tags; pixel labeling can be predicted from image classifiers [32, 34].

The fundamental task of unsupervised semantic segmentation is *grouping*, not *semantics* in terms of *naming*, which is unimportant other than the convenience of tagging segments in the same or different groups. The challenge of unsupervised semantic segmentation is to discover groupings within and across images that capture object- and view-invariance of a category without external supervision, so that (Fig. 1): **1)** A baby's face and body are parts of a whole in the same image; **2)** The whole baby is separated from the rest of the image; **3)** A baby instance is more similar to another baby instance than to a cat instance, despite their different poses, illuminations, and backgrounds.

Several representative approaches have been proposed for tackling this challenge under different assumptions.

- **Visual similarity:** SegSort [26] first partitions each image into segments based on contour cues and then by segment-wise contrastive learning discovers clusters of visually similar segments. However, semantics by visual similarity is far too restrictive: A semantic whole is often made up of visually dissimilar parts. Parts of *body* such as *head* and *torso* look very different; it is not their visual similarity but their spatial adjacency and statistical co-occurrence that bind them together.
- **Spatial stability:** IIC [29] maximizes the mutual information between clusterings from two views of the same image related by a known spatial transformation, enforcing stable clustering while assuming that a fixed number of clusters are equally likely within an image. It works best for coarse and balanced texture segmentation and has major trouble scaling up with the scene complexity.
- **Image-wise feature learning:** [56, 60] train representations on object-centric datasets with multiscale cropping to sharpen the representation within the image. These methods do not work well on scene-centric datasets where an image has more than one dominant semantic class.

Grouping as well as semantics naturally have different levels of granularity: A *hand* is an articulated configuration of a *palm* and five *fingers*, likewise a *person* of a *head*, a *torso*, two *arms*, and two *legs*. Such an inherent grouping hierarchy poses a major challenge: Which level should an unsupervised segmentation method target at and what is the basis for such a determination? Existing methods avoid this ambiguity and treat it as either a factor outside the segmentation modeling, or an aspect of secondary concern.

Our key insight is that the inherent hierarchical organization of visual scenes is not a nuisance for scene parsing, but a universal property that we can exploit and desire for unsupervised segmentation. This idea has previously led to a general image segmenter that handles texture and illusory contours through edges entirely without any explicit characterization of texture or curvilinearity [65]. We now advance the concept to data-driven representation learning: A good representation shall reveal not just a particular level of grouping, but any level of grouping in a consistent and predictable manner across different levels of granularity.

We approach unsupervised semantic segmentation as an unsupervised pixel-wise feature learning problem. Our objective is to best produce a consistent hierarchical segmentation for each image in the entire dataset based entirely on hierarchical clusterings in the feature space (Fig. 1). Specifically, given the pixel-wise feature, we perform hierarchical groupings *within* and *across* images and their transformed versions (i.e.,*views*). In turn, groupings at each level impose a desire on how the feature should be improved to maximize the discrimination among different groups.

Our model has two novel technical components: **1) Multiview cosegmentation** is to not only enforce spatial consistency between segmentations across views, but also bootstrap feature learning from visual similarity and co-occurrences in a simpler clean setting; **2) Clustering transformers** are used to enforce semantic consistency across different levels of the feature grouping hierarchy.

To summarize, our work makes three contributions.

1. **We deliver the first unsupervised hierarchical semantic segmentation** method that can produce parts and wholes in a data-driven manner from an arbitrary collection of images, whether they come from object-centric or scene-centric datasets.
2. **We are the first to embrace the ambiguity of grouping granularity** and exploit the inherent grouping hierarchy of visual scenes to learn a pixel-wise feature representation for unsupervised segmentation. It can thus discover semantics based on not only visual similarity but also statistical co-occurrences.
3. **We outperform existing unsupervised (hierarchical) semantic segmentation methods by a large margin** on not only object-centric but also scene-centric datasets.

2. Related Work

Image segmentation refers to the task of partitioning an image into visually coherent regions. Traditional approaches often consist of two steps: extracting local features and clustering them based on different criteria, *e.g.*, mode-finding [3, 10], or graph partitioning [16, 42, 52, 66, 67].

Hierarchical image segmentation has been supervisedly learned from how humans perceive the organization of an image [2]: While each individual segmentation targets a particular level of grouping, the collection of individual segmentations present the perceptual hierarchy statistically.

A typical choice for representing a hierarchical segmentation is contours: They are first detected to sharply localize region boundaries [25, 63] and can then be removed one by one to reveal coarser segmentations (OWT-UCM [2]).

Such models are trained on individual ground-truth segmentations, hoping that coarse and fine-grained organization would emerge automatically from common and rare contour occurrences respectively in the training data.

In contrast, our model is trained on multi-level segmentations unsupervisedly discovered by feature clustering, and it also operates directly on segments instead of contours.

Semantic segmentation refers to the task of partitioning an image into regions of different semantic classes. Most deep learning models treat segmentation as a spatial extension of image recognition and formulate it as a pixel-wise classification problem. They are often based on Fully Convolutional Networks [7, 36, 40], incorporating information from multiple scales [8, 18, 22–24, 31–33, 35, 45, 53, 64].

SegSort [26] does not formulate segmentation as pixel-wise labeling, but pixel-segment contrastive learning that operates directly on segments delineated by contours. It learns pixel-wise features in a non-parametric way, *with* or *without* segmentation supervision. SPML [32] extends it to unify segmentation with various forms of weak supervision: image-level tags, bounding boxes, scribbles, or points.

Unsupervised semantic segmentation has been modeled by non-parametric methods using statistical features and graphical models [39, 49, 54]. For example, [49] proposes to discover region boundaries by mining the statistical differences of matched patches in coarsely aligned images.

There are roughly three lines of recent unsupervised semantic segmentation methods. **1)** One way is to increase the location sensitivity of the feature learned from images [9, 20, 59, 62], by either adding an additional contrastive loss between pixels based on feature correspondences across views [60], or using stronger augmentation and constrained cropping [51, 56]. **2)** A pixel-level *feature* encoder can be learned directly by maximizing discrimination between pixels based on either contour-induced segments [26] or region hierarchies [68] derived from OWT-UCM [2]. Segmentation is indicated by pixel feature similarity and semantic labels can be inferred from retrieved nearest neighbours in a labeled set. **3)** A pixel-wise *cluster* predictor can be directly learned by maximizing the mutual information between cluster predictions on augmented views of the same instance at corresponding pixels [29, 47].

Our model advances pixel-wise feature learning methods [26, 32, 69]: It contrasts features based on feature-induced hierarchical groupings themselves, and most strikingly, directly outputs consistent hierarchical segmentations.

3. Hierarchical Segment Grouping (HSG)

We approach unsupervised semantic segmentation as an unsupervised pixel-wise feature learning problem (Fig. 2). The basic idea is that, once every pixel is transformed into a point in the feature space, image segmentation becomes a point clustering problem.

Semantic segmentation and feature clustering form a pair of dual processes: **1)** Clustering of feature X defines segmentation G in each image: Pixels with features in the same (different) clusters belong to the same (different) semantic regions. This idea is used to co-segment similar images given handcrafted features [30, 37, 48]. **2)** Segmentation G defines the similarity of feature X: A pixel should be mapped close to its own segment group and far from other segment groups in the feature space. This idea is used to

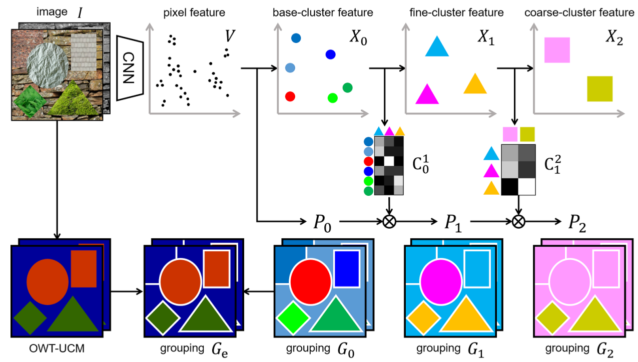

Figure 2. Method overview. We aim to learn a CNN that maps each pixel to a point in the feature space V such that successively derived cluster features X_0, X_1, X_2 produce good and consistent hierarchical pixel groupings G_e, G_1, G_2. Their consistency is enforced through clustering transformers C_l^{l+1}, which dictates how feature clusters at level l map to feature clusters at level $l+1$. Note that G_0 results from clusters of V, and G_e from OWT-UCM edges. P_l is the probabilistic version of G_l, and G_l the winner-take-all binary version of P_l; $P_0 \sim G_0$. For $l \geq 0$, P_{l+1} results from propagating P_l by C_l^{l+1}. Groupings G_e, G_1, G_2 in turn impose desired feature similarity and drive feature learning. We co-segment multiple views of the same image to capture spatial consistency, visual similarity, statistical co-occurences, and semantic hierarchies.

learn the pairwise feature similarity [44] and pixel-wise feature [26, 32] given segmentations.

Our key insight is that a good representation shall reveal not just a particular level of grouping – as past co-segmentation methods have explored, but any level of grouping in a consistent and predictable manner. If we embrace the ambiguity of grouping granularity that all previous methods have avoided and desire the consistency of hierarchical semantic segmentation on the pixel-wise feature, we address not only the shortcoming of cosegmentation, but also provide a joint feature-segmentation learning solution.

Specifically, while there is no supervision available for either feature X or segmentation G, we can desire that: **1)** each segmentation separates features well and **2)** the coarser segmentation defined by next-level feature clusters simply *merges* the current finer segmentation. These strong constraints guide the feature learning towards quality hierarchical segmentations, thereby better capturing semantics.

Our model has two components: **1)** multiview cosegmention to robustify feature clustering against spatial transformation and appearance variations of visual scenes, and **2)** clustering transformers to enforce consistent semantic segmentations across different levels of the feature grouping hierarchy. Both are necessary for mapping pixel features to segmentations, which in turn impose desired pairwise attraction and repulsion on the pixel features.

In the following, we first introduce our contrastive feature learning loss given any grouping G, and then describe how we obtain three kinds of groupings within and across images, and how we evaluate their goodness of grouping and enforce their consistency.

3.1. Pixel-Segment Contrastive Feature Learning

We learn a pixel-wise feature extraction function f as a convolutional neural network (CNN) with parameters θ. It transforms image I to its pixel-wise feature V. Let $\boldsymbol{v}_i$ be the *unit-length* feature vector at pixel i of image I:

$$\boldsymbol{v}_i = f_i(I;\theta), \quad \|\boldsymbol{v}_i\| = 1. \tag{1}$$

Suppose that I is partitioned into segments (Fig. 3). Let $\boldsymbol{u}_s$ be the feature vector for segment s, defined as the (length-normalized) average pixel feature within the segment:

$$\boldsymbol{u}_s \propto \text{mean}\left(\boldsymbol{v}_i : i \text{ in segment } s\right), \quad \|\boldsymbol{u}_s\| = 1 \tag{2}$$

Consider a batch of images and their pixel groupings $\{(I, G)\}$. We want to learn the right feature mapper f such that all the pixels form distinctive clusters in the feature space, each corresponding to a different semantic group.

We follow [26, 32] to formulate desired feature-wise attraction and repulsion *not between pixels*, but *between pixels and segments*. Such contrastive learning across granularity levels reduces computation, improves balance between attraction and repulsion, and is more effective [59].

Our contrastive feature learning loss to minimize is:

$$\mathcal{L}_f(G) = \sum_i -\log \frac{\sum\limits_{s \in G_i^+} \exp \frac{\boldsymbol{v}_i^\top \boldsymbol{u}_s}{T}}{\sum\limits_{s \in G_i^+} \exp \frac{\boldsymbol{v}_i^\top \boldsymbol{u}_s}{T} + \sum\limits_{s \in G_i^-} \exp \frac{\boldsymbol{v}_i^\top \boldsymbol{u}_s}{T}} \tag{3}$$

where T is a temperature hyper-parameter that controls the concentration level of the feature distribution. Ideally, $\boldsymbol{v}_i$ should be attracted to segments in the positive set G_i^+ and repelled by segments in the negative set G_i^-.

Our batch of images consists of several augmented *views* of some training instances. For pixel i in a particular view of image I, G_i^+ includes segments of the same semantic group in any view of image I except i's own segment, in order to achieve within-instance invariance, whereas G_i^- includes segments of different semantic groups in any view of I, *and* segments of training instances other than I, in order to maximize between-instance discrimination [26, 62].

3.2. Consistent Segmentations by View & Hierarchy

From pixel feature V, we compute feature grouping G_0 and cluster feature X_0. Our initial pixel grouping G_e is based on OWT-UCM edges detected in the image. Next-level cluster feature X_{l+1} and grouping G_{l+1} are predicted from G_l with ensured consistency. We use three levels for the sake of illustration (Fig. 3), but our procedure can be repeated for more (coarser) levels.

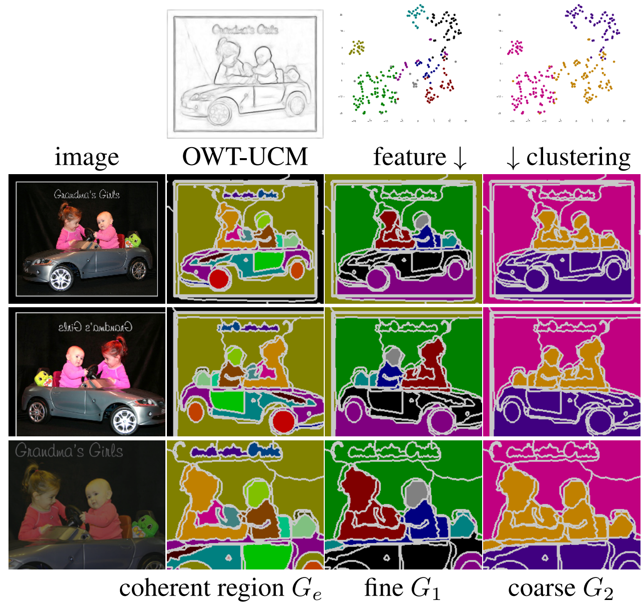

Figure 3. We co-segment multiple views (Column 1) of the same image by OWT-UCM edges (G_e, Column 2) or by feature clustering at fine and coarse levels (G_1, G_2, Columns 3-4). White lines mark the segments derived from pixel feature clustering and OWT-UCM edges. The color of feature points (pixels) mark grouping in the feature space (segmentation in the image) consistently across rows in the same column, per spatial transformations between views. G_2's coarse segmentations simply merge G_1's fine segmentations, their consistency enforced by our clustering transformers. Minimizing $\mathcal{L}_f(G_e), \mathcal{L}_f(G_1), \mathcal{L}_f(G_2)$ ensures respectively that our learned feature is grounded in low-level coherence, yet with view invariance, and capable of capturing semantics at multiple levels and producing hierarchical segmentations.

Base cluster feature X_0 and grouping G_0, G_e. We segment each view of I by clustering pixel features, resulting in base grouping G_0 and cluster (centroid) feature X_0 (Fig. 2).

During training but *not* testing, we segment image I into a fixed number of coherent regions according to its OWT-UCM edges [14], based on which we split each G_0 region to obtain edge-conforming *segments* [26] marked by white lines in Fig. 3. For training, we obtain pixel grouping G_e by inferring the coherent region segmentation according to how each view is spatially transformed from I.

Minimizing $\mathcal{L}_f(G_e)$ encourages the feature to be similar not only for different pixels of similar appearances in the image, but also for corresponding pixels of different appearances across views of I. The former grounds the feature f at respecting low-level appearance coherence, whereas the latter develops view invariance in the feature.

Next-level cluster feature X_{l+1} and grouping G_{l+1}. Now we have grouping G_0 in the feature space of V, and for each cluster, we obtain its centroid feature in X_0. We model how cluster feature X_l maps to cluster feature X_{l+1}, which corresponds to how segmentation at level l maps to segmentation at level $l+1$ in the image.

We adopt a probabilistic framework, where any feature point $\boldsymbol{x}$ has a (soft assignment) probability belonging to a group determined by its cluster centroid. Let $P_l(a)$ be the probability of $\boldsymbol{x}$ in group a at level l:

$$P_l(a) = \text{Prob}(G_l = a \,|\, \boldsymbol{x}). \quad (4)$$

To ensure that feature points in the same group remain together at the next level, we introduce group transition probability $C_l^{l+1}(a, b)$, the transition probability from group a at level l to group b at level $l+1$:

$$C_l^{l+1}(a, b) = \text{Prob}(G_{l+1} = b \,|\, G_l = a). \quad (5)$$

Per the Bayesian rule, we have:

$$P_{l+1}(b) = \sum_a P_l(a) \cdot C_l^{l+1}(a, b). \quad (6)$$

Writing P_l as a row vector, we can derive the soft group assignment P_{l+1} for cluster feature X_0 at level $l+1$:

$$P_{l+1} = P_l \times C_l^{l+1} = P_0 \times C_0^1 \times C_1^2 \times \cdots \times C_l^{l+1}. \quad (7)$$

Clustering Transformers. C_l^{l+1} is defined on multiview cosegmentation of each instance. We learn a function, in terms of a transformer [5], to naturally capture feature group transitions for all the training instances. It enables more consistent grouping compared to non-parametric clustering methods such as KMeans, NCut [58], and FINCH [50].

Our clustering transformer from level l to $l+1$ maps group centroid feature X_l to the next-level group centroid feature X_{l+1}, and simultaneously outputs the group transition probability C_l^{l+1} (Fig. 4).

Consistent feature groupings. At level $l=0$, P_0 has binary values, indicating hard grouping G_0. For next level l, we compute P_{l+1} by propagating P_l with our clustering transformer C_l^{l+1}, which also outputs X_{l+1}. We obtain G_{l+1} by binarizing P_{l+1} with winner-take-all. By decreasing the number of groups as l increases, we obtain consistent fine to coarse segmentations G_1, G_2 (Fig. 2).

Minimizing $\mathcal{L}_f(G_1)$ and $\mathcal{L}_f(G_2)$ encourages the feature f to capture semantics at multiple levels and produce consistent hierarchical segmentations (Fig. 3).

3.3. Goodness of Grouping

While clustering transformers ensure grouping consistency across levels, we still need to drive feature learning towards good segmentations. We follow [55] and supervise our transformer with modularity maximization [46] and collapse regularization. The former seeks a partition that results higher (lower) in-cluster (out-cluster) similarity than the total expectation, whereas the latter encourages

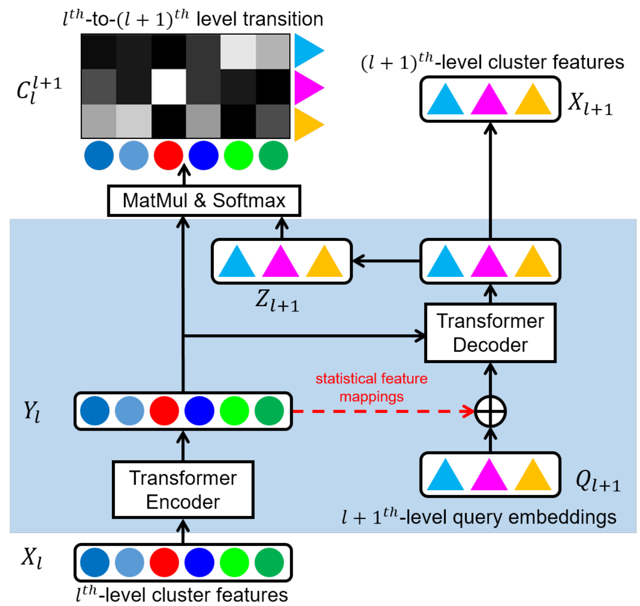

Figure 4. Our clustering transformer enforces grouping consistency across levels by mapping feature X_l to X_{l+1} with feature transition C_l^{l+1}. X_{l+1} and C_l^{l+1} are learned simultaneously. Shown here for level $l=0$ in Fig. 2, the transformer encoder also takes learnable inputs from query embeddings Q_l and outputs contextualized feature Y_l. The transformer decoder outputs X_{l+1} and additionally projected feature Z_{l+1}. The transition is predicted as: $C_l^{l+1} = \text{softmax}\left(\frac{1}{\sqrt{m}} Y_l^\top Z_{l+1}\right)$; m is the feature dimension. Statistical feature mapping: Calculate Y_l's mean and std, transform them by fc layers, and add to Q_l for instance adaptation.

partitions of equal sizes. We additionally maximize the separation between cluster centroids.

We first build a sparsified graph based on pairwise feature similarity for X_0. Let e be the number of edges in this graph, n_l the number of centroids in X_l, A the $n_0 \times n_0$ connection matrix for edges, D the $n_0 \times 1$ degree vector of A, M_l the $n_0 \times n_l$ soft assignment matrix where each row is P_l for a centroid of X_0, and $\boldsymbol{z}_{l,k}$ the normalized k-th feature of Z_l in Fig. 4. Our goodness of grouping loss is:

$$\mathcal{L}_g = \sum_{l \geq 1} \underbrace{\frac{-1}{2e} \text{trace}(M_l^\top (A - \frac{1}{2e} DD^\top) M_l)}_{\text{maximize modularity}} + \underbrace{\frac{\sqrt{n_l}}{n_0} \|1^\top M_l\|_F - 1}_{\text{collapse regularization}}$$
$$+ \underbrace{\frac{1}{n_l} \sum_k -\log \frac{\exp(\boldsymbol{z}_{l,k}^\top \boldsymbol{z}_{l,k})}{\sum_j \exp(\boldsymbol{z}_{l,k}^\top \boldsymbol{z}_{l,j})}}_{\text{maximize centroid separation}} \quad (8)$$

3.4. Model Overview: Training and Testing

Our model (Fig. 5) is trained with the contrastive feature learning losses given edge-based grouping G_e and multi-level feature-based grouping G_l, and the goodness of group-

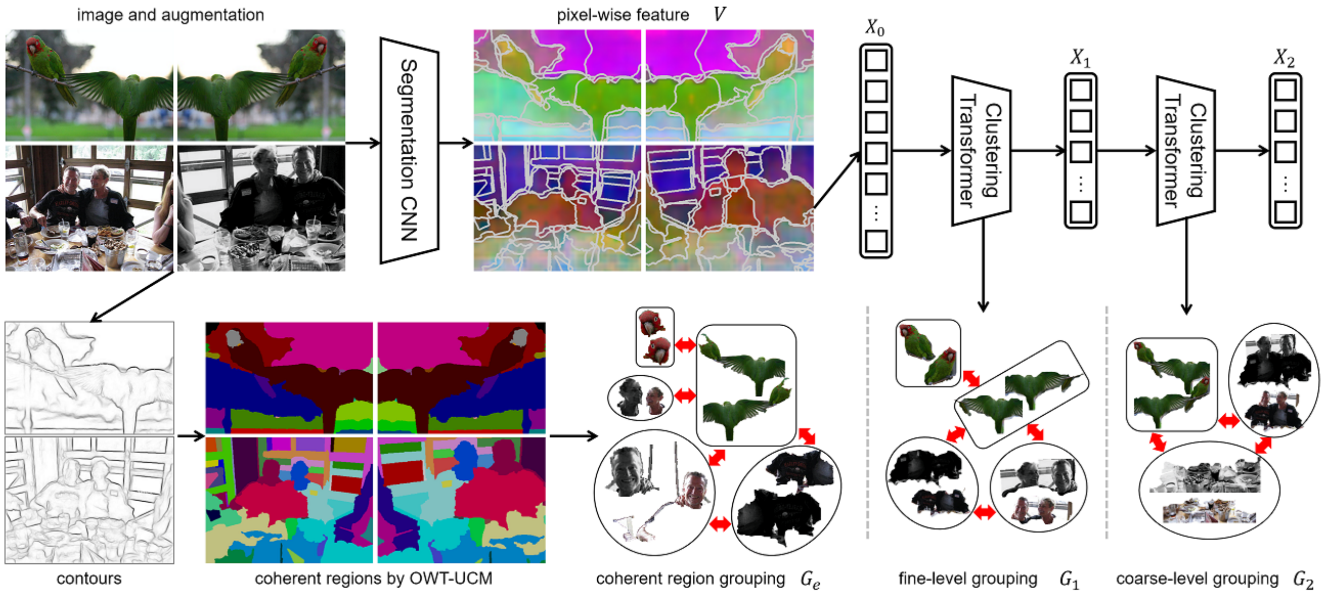

Figure 5. Our model consists of two essential components: **1)** multiview cosegmentation and **2)** hierarchical grouping. We first produces pixel-wise feature V, from which we cluster to get base cluster feature X_0 and grouping G_0. Each G_0 region is split w.r.t coherent regions derived by OWT-UCM procedure, which is marked by the white lines. We create three groupings–G_e, G_1 and G_2 in multiview cosegmentation fashion. We obtain G_e by inferring the coherent region segmentation according to how each view is spatially transformed from the original image. Starting with input X_0 of an image and its augmented views, we conduct feature clustering to merge G_0 into G_1, and then, G_1 into G_2. Based on G_e, G_1 and G_2, we formulate a pixel-to-segment contrastive loss for each grouping. Our HSG learns to generate discriminative representations and consistent hierarchical segmentations for the input images.

ing loss, weighted by λ_E, λ_F, and λ_G respectively:

$$\mathcal{L}(f) = \lambda_E \mathcal{L}_f(G_e) + \lambda_F \sum_{l \geq 1} \mathcal{L}_f(G_l) + \lambda_G L_g. \tag{9}$$

For testing, the same pipeline with the pixel feature CNN and clustering transformers predicts hierarchical segmentations $\{G_l\}$. To benchmark segmentation performance given a labeled set, We follow [26] and predict the labels using k-nearest neighbor search for each segment feature.

4. Experiments

We benchmark our model on two tasks: unsupervised semantic segmentation and hierarchical image segmentation, the first on five major object- and scene-centric datasets and the second on Pascal VOC. We conduct ablation study to understand the contributions of our model components.

We adopt FCN-ResNet50 as the common backbone architecture. The FCN head consists of 1×1 convolution, $\mathrm{BatchNorm}$, ReLU, and 1×1 convolution. Specifically, we follow DeepLabv3 [8] to set up the dilation and strides in ResNet50. We set $\mathrm{Multi\_Grid}$ to $(1, 2, 4)$ in $\mathrm{res5}$. The $\mathrm{output\_stride}$ is set to 16 and 8 during training and testing. We do not use any pre-trained models, but train our models from scratch on each dataset. Ground-truth annotations are not for training but only for testing and evaluation's sake.

Pascal VOC 2012 [15] is a generic semantic segmentation dataset of 20 object category and a background class. It consists of $1,464$ and $1,449$ images for training and validation. We follow [7] to augment the training data with additional annotations [19], resulting in $10,582$ training images. Following [56], we do not train but only inference on VOC.

MSCOCO [38] is a complex scene parsing dataset with 80 object categories. Objects are embedded in more complex scenes, with more objects per image than Pascal (7.3 vs. 2.3). Following [56, 60], we use *train2017* split ($118,287$ images) for training and test on the VOC validation set.

Cityscapes [11] is an urban street scene parsing dataset, with 19 stuff and object categories. Unlike MSCOCO and VOC where classes are split by scene context, Cityscapes contains similar street scenes covering almost all 19 categories. The train/test split is $2,975/500$.

KITTI-STEP [61] is a video dataset for urban scene understanding, instance detection and object tracking. It has pixel-wise labels of the same 19 categories as Cityscapes. There are 12 and 9 video sequences for training and validation, or $5,027$ and $2,981$ frames.

COCO-stuff [4] is a scene texture segmentation dataset, a subset of MSCOCO. As [29, 47], we use 15 coarse *stuff* categories and reduce the dataset to 52K images with at least 75% stuff pixels. The train/test split is $49,629/2,175$.

Potsdam [17] is a dataset for aerial scene parsing. The raw

6000×6000 image is divided into 8550 RGBIR 200×200 patches. There are 6 categories (*roads, cars, vegetation, trees, buildings, clutter*). The train/test split is $7,695/855$.

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**

Table 1. Our method delivers better performance on different types of datasets. The results are reported on VOC, KITTI-STEP and Cityscapes val set, using IoU and pixel accuracy metrics. In VOC, object categories are separated according to image scenes. In Cityscapes and KITTI-STEP, images all come from urban street scene and thus contain mostly the same set of categories. Instance-discrimination methods apply image-wise contrastive loss, and learn less optimally on Cityscapes and KITTI-STEP, as image scenes are similar. Our HSG instead learns to discriminate regions at different scales and performs well on both types of datasets.

	COCO-stuff		Potsdam	
Method	mIoU	Acc.	mIoU	Acc.
DeepCluster 2018 [6]	-	19.9	-	29.2
Doersch 2015 [13]	-	23.1	-	37.2
Isola 2016 [28]	-	24.3	-	44.9
IIC [29]	-	27.7	-	45.4
AC [47]	-	30.8	-	49.3
SegSort [26]	16.4	49.9	35.0	59.0
Our HSG	**23.8**	**57.6**	**43.8**	**67.4**

Table 2. Our method outperforms baselines on both stuff region and aerial scene parsing datasets. The results are reported on COCO-stuff and Potsdam test set, using IoU and pixel accuracy metrics. We evaluate our model using nearest neighbor search. Our HSG achieves superior performance.

λ_E	λ_G	λ_F	single-view	multi-view
✓	-	-	13.0	40.9
✓	✓	-	13.8	41.7
✓	✓	✓	14.0	41.9

Table 3. Regularizing with our goodness of grouping loss and pixel-to-segment contrastive losses improves learned features. The results are reported over VOC val set, using IoU metric. Our resulted pixel features encode better semantic information.

Method	KMeans	NCut [58]	FINCH [50]	Our Transfomer
mIoU	41.2	41.3	40.6	**41.9**

Table 4. Our hierarchical clustering transformer follows semantics closer than other non-parametric clustering algorithms. The results are reported on VOC val set with IoU metric. Our learned representations achieve better unsupervised semantic segmentation.

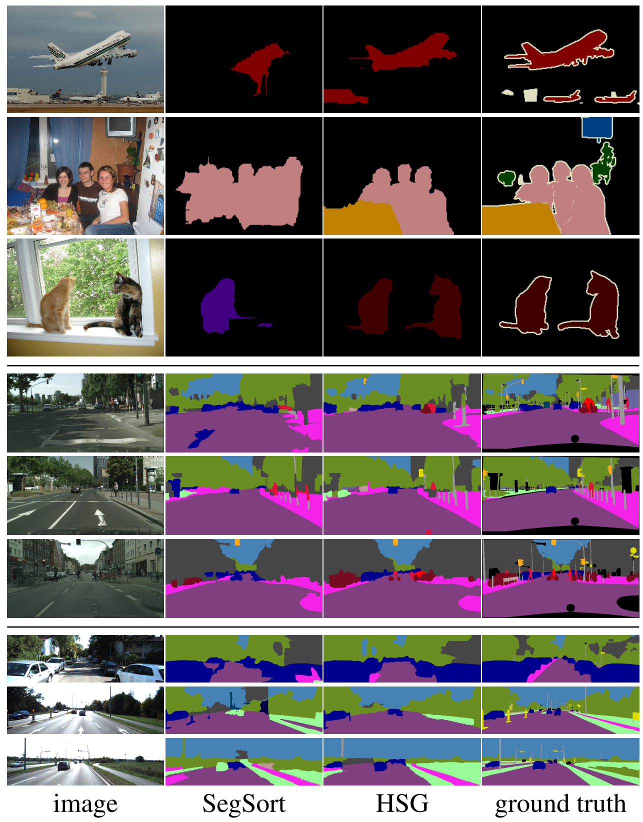

Figure 6. Our framework performs better on different types of datasets. From top to bottom every three rows are visual results from VOC, Cityscapes and KITTI-STEP dataset. The results are predicted via segment retrievals. Our pixel-wise features encode more precise semantic information than baselines.

Results on unsupervised semantic segmentation. All the models are trained from scratch and evaluated by IoU and pixel accuracy. For VOC, we follow baselines [56] to train on MSCOCO. Table. 1 shows that our method outperforms baselines by 6.8%, 7.9% and 2.5% in mIoU on VOC, Cityscapes, and KITTI-STEP validation sets respectively.

Note that methods relying on image-wise instance discrimination do not work well on Cityscapes and KITTI-STEP. Both datasets have urban street scenes with similar categories in each image. Our method can still discover semantics by discriminating regions among these images.

For texture segmentation on COCO-stuff and Potsdam,

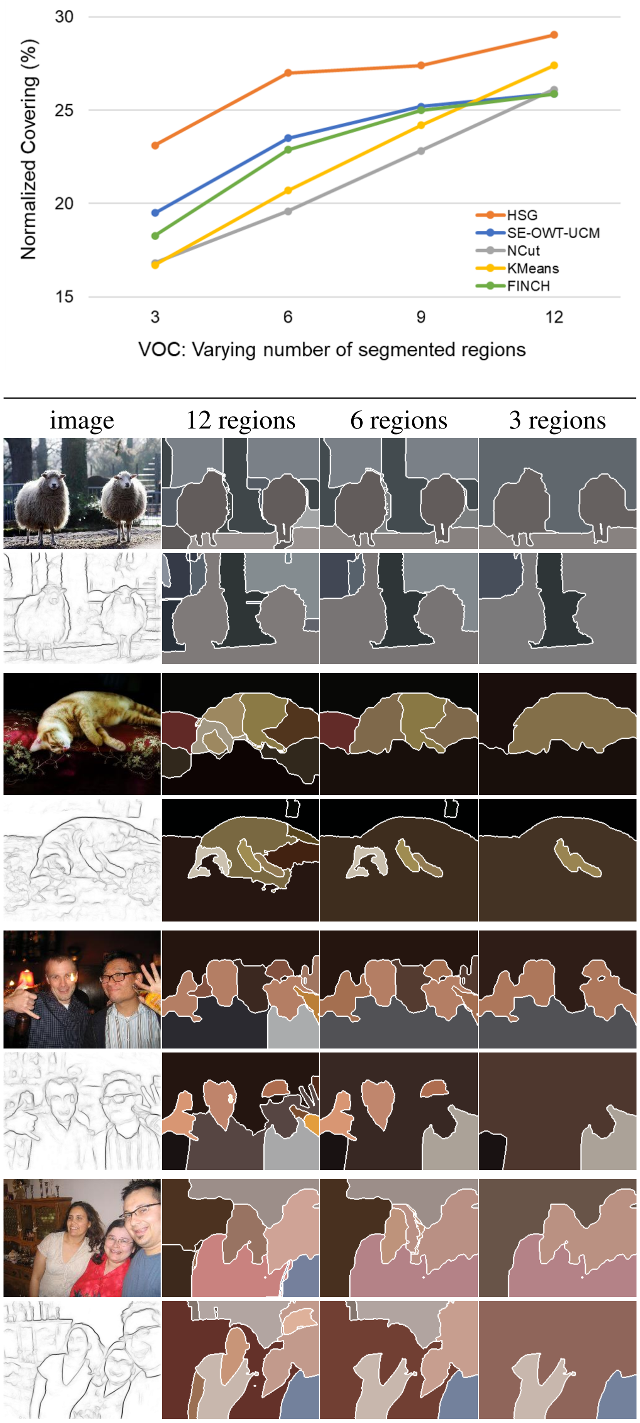

Figure 7. Our clustering transformers capture semantics at different levels of granularity. **Top:** We compare to other clustering algorithms on VOC val set, using *Normalized Foreground Coverings* as metric. We exclude background regions for evaluation. Our HSG overlaps with ground truths more accurately. **Bottom:** We present visual results to compare our hierarchical segmentation (top row) with SE [14]-OWT-UCM procedure (bottom row). We also show the detected edges at the leftmost figure in the bottom row. Each image is segmented into 12, 6, 3 regions. Our method reveals low-to-high level of semantics more consistently.

Tab. 2 shows that our method achieves huge gains, $+26.8\%$ and $+18.1\%$ over IIC [29] and AC [47] respectively.

Results on hierarchical segmentation. We benchmark hierarchical segmentation with respect to ground-truth segmentation. We evaluate the overlapping of regions between predicted segmentations and ground truth within each image, known as *Segmentation Covering* [2]. However, such a metric scores performance with the number of pixels within each segment, and is thus easily biased towards large regions. For object-centric dataset VOC, a trivial all-foreground mask would rank high by the Covering metric.

We propose a *Normalized Foreground Covering* metric, by focusing on the foreground region and the overlap ratio instead of the overlap pixel count. To measure the average foreground region overlap ratio of a ground-truth segmentation S by a predicted segmentation S', we define:

$$\text{NFCovering}(S' \to S_{fg}) = \frac{1}{|S_{fg}|} \sum_{R \in S_{fg}} \max_{R' \in S'} \frac{|R \cap R'|}{|R \cup R'|} \quad (10)$$

where S_{fg} denotes the set of ground-truth foreground regions. Given a hierarchical segmentation, we report NF-Covering at each level in the hierarchy. Fig. 7 shows that our clustering transformers produce segmentations better aligned with the ground-truth foreground at every level.

Visualization. Fig. 6 shows sample semantic segmentations on VOC (trained on MSCOCO), Cityscapes and KITTI-STEP. Compared to SegSort [26], our method retrieves same-category segments more accurately. For larger objects or stuff categories, such as *airplane* or *road*, our results are more consistent within the region. Our segmentations are also better at respecting object boundaries.

We also compare our hierarchical segmentations with SE [14]-OWT-UCM, an alternative based entirely on low-level cues. Fig. 7 bottom shows that, when partitioning an image into 12, 6 and 3 regions, our segmentations follow the semantic hierarchy more closely.

Ablation study. Tab. 3 shows that our model improves consistently by adding the feature learning loss based on hierarchical groupings and the goodness of grouping loss. It also shows that multiview cosegmentation significantly improves the performance over a single image.

Tab. 4 shows that our clustering transformers provide better regularization with hierarchical groupings than alternative non-parametric clustering methods.

Summary. We deliver the first unsupervised hierarchical semantic segmentation method based on multiview cosegmentation and clustering transformers. Our unsupervised segmentation outperforms baselines on major object- and scene-centric benchmarks, and our hierarchical segmentation discovers semantics far more accurately.

Acknowledgements. This work was supported, in part, by Berkeley Deep Drive, Berkeley AI Research Commons with Facebook, NSF 2131111, and a Bosch research gift.

References

[1] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 12, 14

[2] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):898–916, 2010. 2, 3, 8, 15

[3] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005. 2

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 6

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 5, 15

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 7

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 3, 6

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3, 6

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 3

[10] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *PAMI*, 2002. 2

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 13, 17

[13] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 7

[14] Piotr Dollár and C Lawrence Zitnick. Fast edge detection using structured forests. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1558–1570, 2014. 4, 8, 17

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 6

[16] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 2004. 2

[17] Markus Gerke. Use of the stair vision library within the isprs 2d semantic labeling benchmark. 2014. 6

[18] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009. 3

[19] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 6

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 3, 7

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 17

[22] Xuming He, Richard S Zemel, and MA Carreira-Perpindn. Multiscale conditional random fields for image labeling. In *CVPR*, 2004. 3

[23] Jyh-Jing Hwang, Tsung-Wei Ke, Jianbo Shi, and Stella X Yu. Adversarial structure matching for structured prediction tasks. In *CVPR*, 2019.

[24] Jyh-Jing Hwang, Tsung-Wei Ke, and Stella X Yu. Contextual image parsing via panoptic segment sorting. In *Multimedia Understanding with Less Labeling on Multimedia Understanding with Less Labeling*, pages 27–36. 2021. 3

[25] Jyh-Jing Hwang and Tyng-Luh Liu. Pixel-wise deep learning for contour detection. *arXiv preprint arXiv:1504.01989*, 2015. 2

[26] Jyh-Jing Hwang, Stella X Yu, Jianbo Shi, Maxwell D Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. Segsort: Segmentation by discriminative sorting of segments. In *ICCV*, 2019. 1, 2, 3, 4, 6, 7, 8, 13, 14, 15

[27] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Crisp boundary detection using pointwise mutual information. In *European Conference on Computer Vision*, pages 799–814. Springer, 2014. 17

[28] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Learning visual groups from co-occurrences in space and time. *arXiv preprint arXiv:1511.06811*, 2015. 7

[29] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019. 2, 3, 6, 7, 8

[30] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, pages 1943–1950. IEEE, 2010. 3

[31] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X. Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, 2018. 3

[32] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment con-

trastive learning. In *International Conference on Learning Representations*, 2021. 1, 3, 4

[33] Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009. 3

[34] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016. 1

[35] Lubor Ladicky, Christopher Russell, Pushmeet Kohli, and Philip HS Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 3

[36] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989. 3

[37] Yong Jae Lee and Kristen Grauman. Collect-cut: Segmentation with top-down cues discovered in multi-object images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3185–3192. IEEE, 2010. 3

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[39] Ce Liu, Jenny Yuen, and Antonio Torralba. Nonparametric scene parsing via label transfer. *PAMI*, 2011. 3

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[41] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 13, 17

[42] Jitendra Malik, Serge Belongie, Thomas Leung, and Jianbo Shi. Contour and texture analysis for image segmentation. *IJCV*, 2001. 2

[43] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001. 17

[44] Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems 13*, 2000. 3

[45] Mohammadreza Mostajabi, Payman Yadollahpour, and Gregory Shakhnarovich. Feedforward semantic segmentation with zoom-out features. In *CVPR*, 2015. 3

[46] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006. 5

[47] Yassine Ouali, Celine Hudelot, and Myriam Tami. Autoregressive unsupervised image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 3, 6, 7, 8

[48] Carsten Rother, Tom Minka, Andrew Blake, and Vladimir Kolmogorov. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into mrfs. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 993–1000. IEEE, 2006. 3

[49] Bryan Russell, Alyosha Efros, Josef Sivic, Bill Freeman, and Andrew Zisserman. Segmenting scenes by matching image composites. In *NIPS*, 2009. 3

[50] Saquib Sarfraz, Vivek Sharma, and Rainer Stiefelhagen. Efficient parameter-free clustering using first neighbor relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2019. 5, 7

[51] Ramprasaath R Selvaraju, Karan Desai, Justin Johnson, and Nikhil Naik. Casting your model: Learning to localize improves self-supervised representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11058–11067, 2021. 3

[52] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000. 2

[53] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009. 3

[54] Joseph Tighe and Svetlana Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *ECCV*, 2010. 3

[55] Anton Tsitsulin, John Palowitch, Bryan Perozzi, and Emmanuel Müller. Graph clustering with graph neural networks. *arXiv preprint arXiv:2006.16904*, 2020. 5

[56] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Revisiting contrastive methods for unsupervised learning of visual representations. *arXiv preprint arXiv:2106.05967*, 2021. 1, 2, 3, 6, 7

[57] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. *arXiv preprint arXiv:2102.06191*, 2021. 13, 14

[58] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007. 5, 7

[59] Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12586–12595, 2021. 1, 3, 4

[60] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3024–3033, 2021. 2, 3, 6, 7

[61] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 6

[62] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018. 1, 3, 4

[63] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 2

[64] Jian Yao, Sanja Fidler, and Raquel Urtasun. Describing the

scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 3

[65] Stella X. Yu. Segmentation induced by scale invariance. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[66] Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, 2003. 2

[67] Stella X Yu and Jianbo Shi. Segmentation given partial grouping constraints. *PAMI*, 2004. 2

[68] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. *Advances in Neural Information Processing Systems*, 33, 2020. 3, 17

[69] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Dazhi Cheng, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, 2018. 3

[70] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 13, 17