NEURAL INFORMATION PROCESSING SYSTEMS

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure

Simone Rossetti Fiora Pirri

DIAG
Dipartimento di Ingegneria informatica, automatica e gestionale Antonio Ruberti

SAPIENZA
UNIVERSITÀ DI ROMA

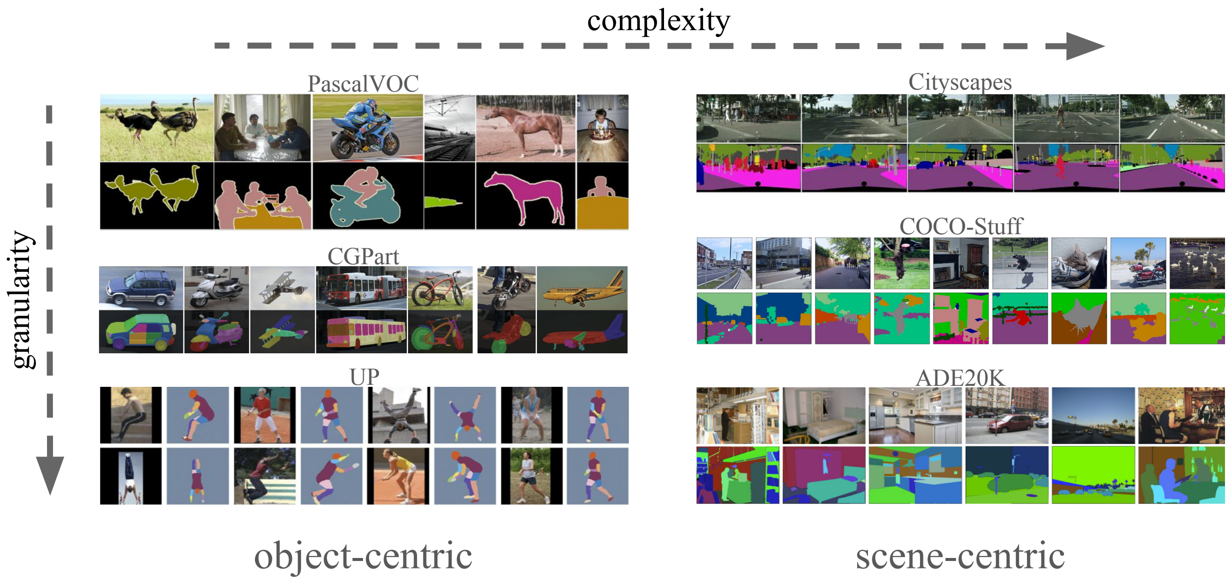
DeepPlants

Motivation and Problem Statement

Challenge: Achieving unsupervised semantic segmentation that can parse complex image structures without external labels or dataset-specific priors.

Key Issue: Existing methods struggle with adapting to dataset-specific varying levels of granularity and often rely on assumptions that limit their generalizability.

Motivation and Problem Statement (contd.)



semantic segmentation datasets differs in semantic granularity and data complexity

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Motivation and Problem Statement (contd.)

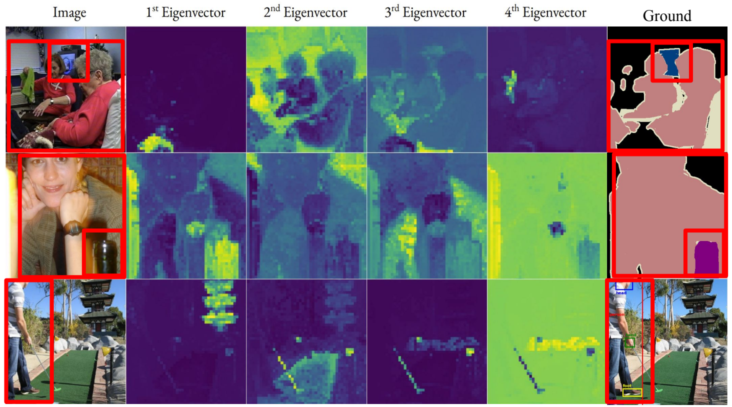
- Non-foreground is missed while some foreground objects are merged

Unsupervised Semantic Segmentation by Contrasting Object Mask Proposals, Van Gansbeke W. et al., ICCV 2021.

Motivation and Problem Statement (contd.)

- Some parts are missed and some other are merged

K-parts foreground extraction



Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization, Melas-Kyriazi L. et al., CVPR 2022.

Motivation and Problem Statement (contd.)

- Same objects with hidden parts are mistakenly divided into more categories

parts over clustering | community detection | ground truth

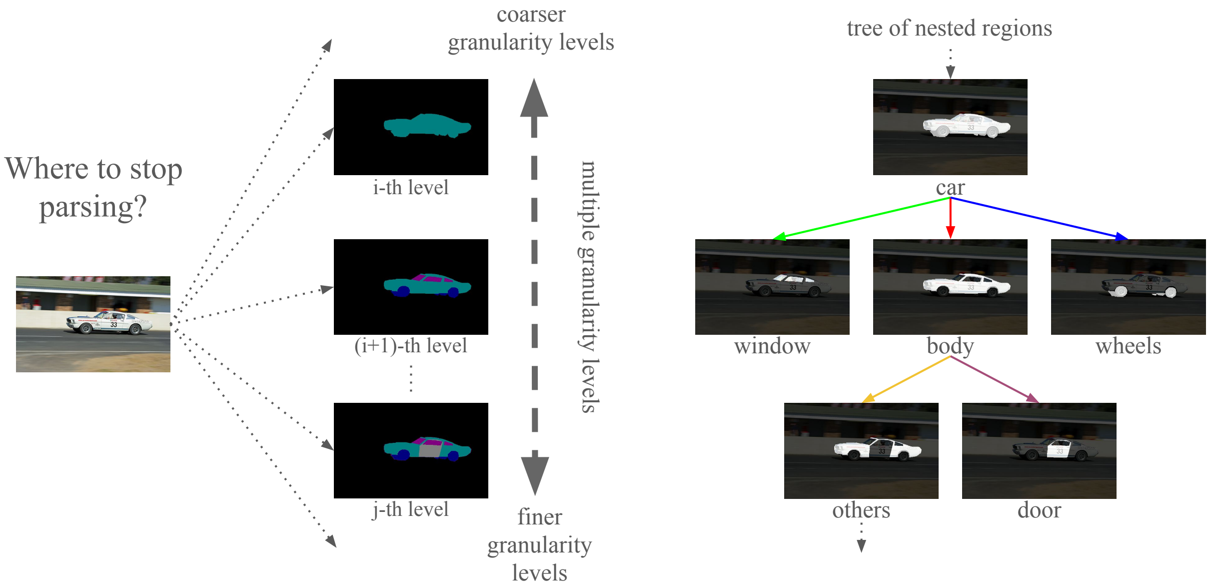
Self-Supervised Learning of Object Parts for Semantic Segmentation, Ziegler A. and Asano Y. M., CVPR 2022.

Motivation and Problem Statement (contd.)

- Objects that share many parts are mistakenly merged to one category

Self-Supervised Learning of Object Parts for Semantic Segmentation, Ziegler A. and Asano Y. M., CVPR 2022.

Motivation and Problem Statement (contd.)



semantics naturally has different levels of granularity

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Main Contributions

1. **Innovative Clustering Method:** Introduction of recursive deep spectral clustering that discerns semantic regions across *multiple granularity levels* without *predefined hierarchies*.
2. **New Evaluation Metrics:** Proposal of *Normalized Multigranular Covering* (NMCovering) and *Normalized Hierarchical Covering* (NHCovering) to benchmark segmentation quality and hierarchy consistency.
3. **Broad Applicability:** Demonstrates versatility when integrated into different self-supervised models, performing well across diverse datasets.

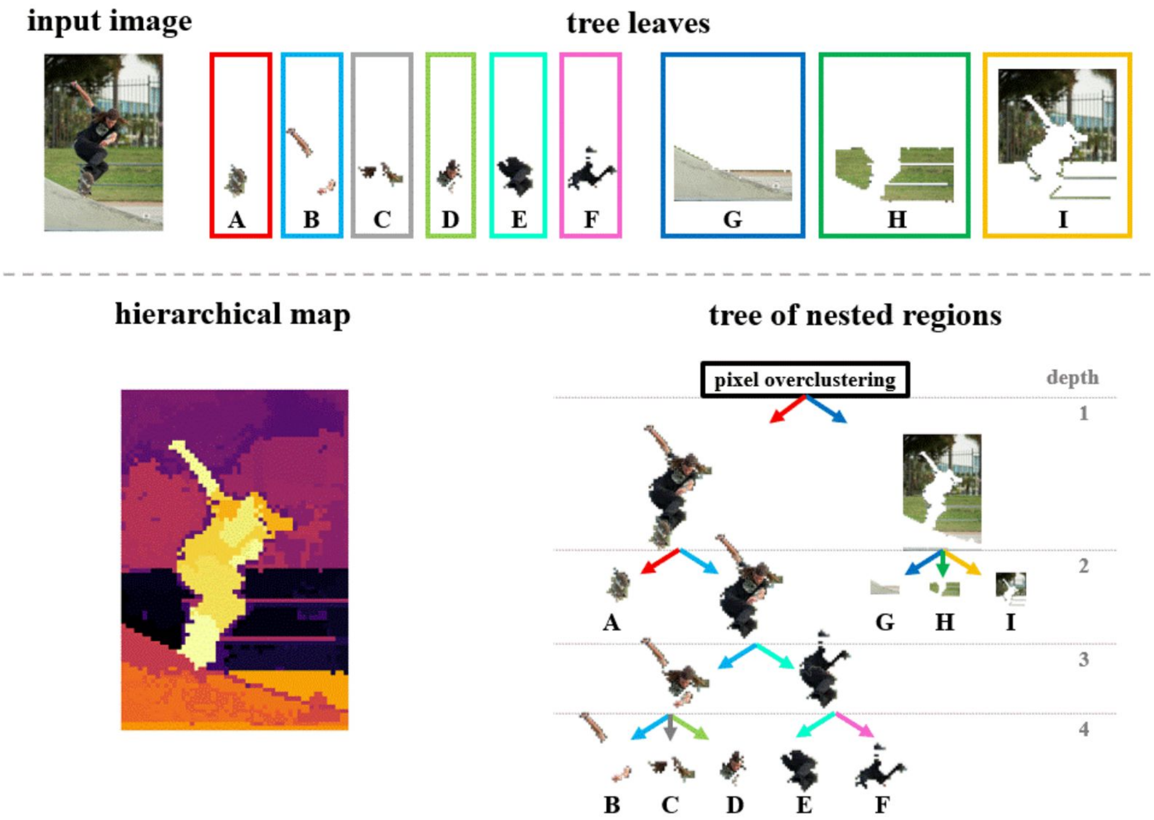
Method Overview

Graph Representation: We represent images as weighted undirected graphs using feature vectors from self-supervised models (e.g., DINO, CLIP) as nodes, with edge weights based on cosine similarity.

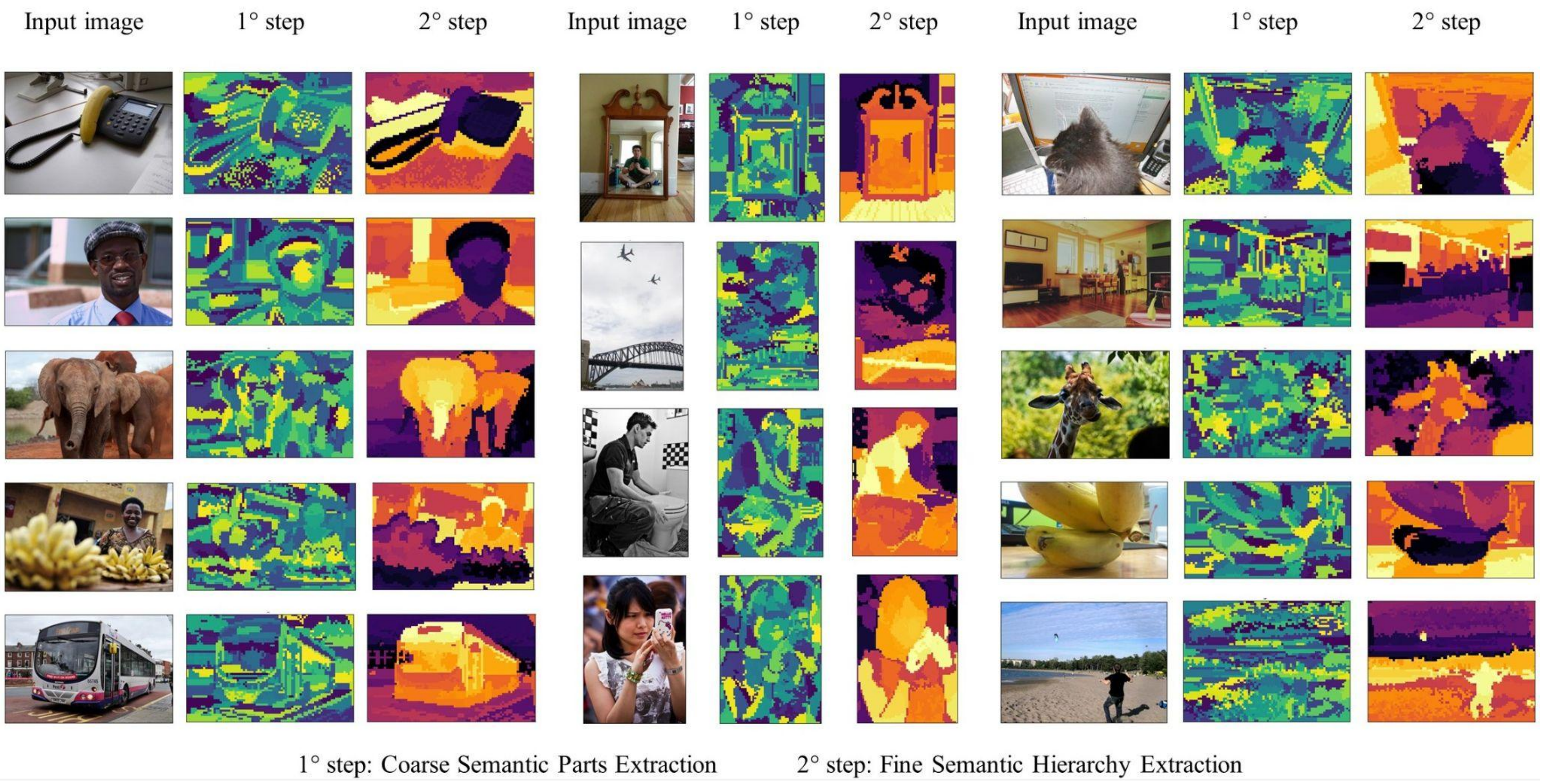
Recursive Clustering Strategy:

- Begins with coarse segments and recursively refines to finer details.
- Utilizes spectral clustering guided by perturbation theory to handle semantic inconsistencies.

Key Concept: The adjacency matrix’s spectral properties are leveraged to partition graphs into semantically consistent subgraphs, refining the image into a *tree of nested regions*.



Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.



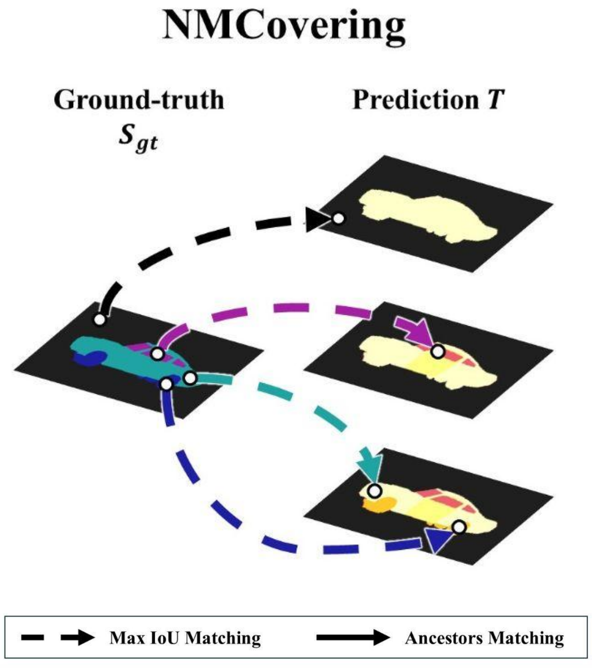
Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Evaluation of a granularity-agnostic grouping

Normalised Multigranular Covering (NMCovering):

- Image-level Jaccard's Index between:
  - Single-level granularity of ground-truth,
  - Multi-level granularity of prediction.
- Semantic lineage not available.

$$\mathrm{NMCovering}(T \to S_{gt}) := \frac{1}{|S_{gt}|} \sum_{R \in S_{gt}} \max_{R' \in T} \frac{|R \cap R'|}{|R \cup R'|}$$

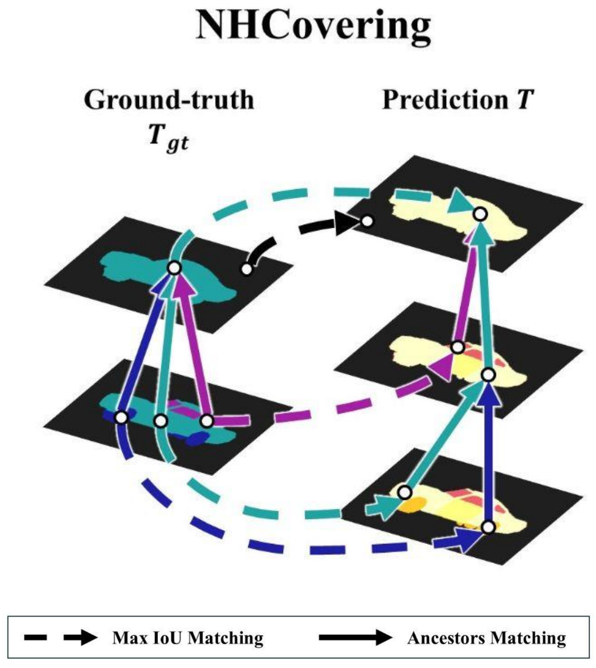


Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Evaluation of a hierarchy-agnostic grouping

Normalised Hierarchical Covering (NHCovering):

- Image-level Jaccard's Index between:
  - Multi-level granularity of ground-truth,
  - Multi-level granularity of prediction.
- Evaluate hierarchical coherence with no labels.



$$\mathrm{NHCovering}(T \to T_{gt}) := \frac{1}{|T_{gt}|} \sum_{R \in T_{gt}} \max_{R' \in T} \frac{|R \cap R'|}{|R \cup R'|} \cdot \frac{|\beta(R,T) \cap \pi(R')|}{|\pi(R)|}$$

$$\text{where } \beta(R,T) := \bigcup_{P \in \pi(R)} \arg\max_{P' \in T} \frac{|P \cap P'|}{|P \cup P'|}.$$

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Experimental Validation

Datasets: Tested on diverse image sets including *PascalVOC2012*, *COCO-Stuff*, *Cityscapes*, and *PartImageNet*, covering both object- and scene-centric challenges.

Results:

- Demonstrated high performance with superior NMCovering and NHCovering scores, indicating effective multi-granular segmentation.
- Outperformed state-of-the-art unsupervised segmentation methods, showing adaptability and finer semantic decomposition.

Main Results

- Object- and scene-centric results;
- Whole- and part-centric results;
- Supervision strategy comparisons;
- Hierarchical clustering comparisons;

Table 1: **Granularity-agnostic.** Evaluation of our algorithm on different datasets using a maximum overlap heuristic for category matching.

Dataset	mIoU	pAcc	mAcc	fIoU	NMCovering $(T \rightarrow S_{gt})$
object-centric					
PascalVOC2012	78.1	82.6	91.2	78.1	75.4
MSCOCO	55.7	93.1	85.0	78.8	49.6
scene-centric					
COCO-Stuff	58.7	81.1	80.3	67.3	42.1
Cityscapes	48.8	82.8	76.1	68.8	44.8
KITTI-STEP	51.2	79.8	76.5	65.7	48.4
Mapillary Vistas	47.6	78.9	72.1	66.1	42.7
Potsdam	58.9	83.4	83.2	65.0	56.3

Table 2: **Hierarchy-agnostic.** Evaluation of our algorithm on different datasets using a maximum overlap heuristic for category matching.

Dataset	mIoU	pAcc	NMCovering $(T \rightarrow T_{gt})$	NHCovering
whole-centric				
COCO-Stuff	59.5	75.1	53.5	42.9
Cityscapes	53.7	78.8	51.1	43.8
KITTI-STEP	58.3	79.6	54.2	46.5
Mapillary Vistas				
part-centric				
Pascal-Part	25.8	80.0	39.5	38.8
Part-Imagenet	55.4	79.5	65.8	65.2
Part-Imagenet-158	59.5	82.6	67.8	63.1

Table 3: **Semantic segmentation.** Comparison on PascalVOC2012 *val*. Ours match unsupervised masks to best overlapping classes.

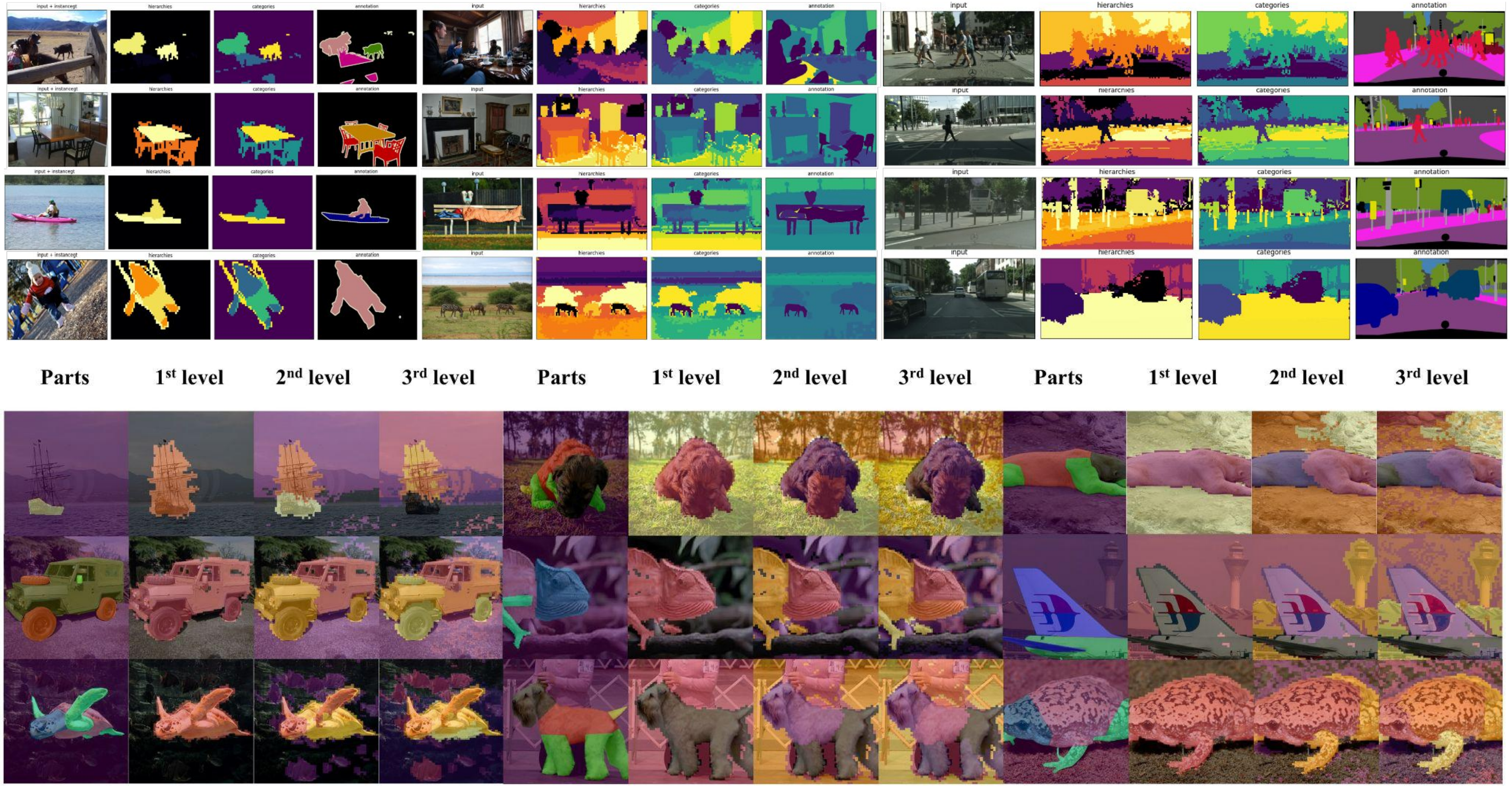
Method	**Backbone**	mIoU VOC12	mIoU MSCOCO
fully-supervised			
DeepLab-CRF [12]	ResNet-101	77.7	-
DeepLab-CRF [12]	VGG-16	-	**43.6** [10]
DeepLabV3-JFT [13]	ResNet-101	**82.7**	-
weakly-supervised			
ViT-PCM [71]	ViT-B16	69.3	45.0
L2G [42]	ResNet-38	72.0	44.2
WeakTr [95]	DeiT-S	**74.0**	**50.3**
un-supervised			
Melas-Kyriazi et al. [58]	ViT-S16	37.2	-
Leopart [96]	ViT-S16	41.7	49.2
HSG [44]	ResNet-50	41.9	-
Zhang et al. [94]	ResNet-50	43.5	-
MaskDistill [79]	ResNet-50	48.9	-
Ours w/o CRF	ViT-S8	76.2 $\pm$ .9	52.1 $\pm$ .6
Ours w CRF	ViT-B14	**80.3** $\pm$ 1.1	**56.5** $\pm$ .9

Table 4: **Boundary potential methods.** All methods match unsupervised tree segments to best overlapping classes.

PascalVOC2012	mIoU	pAcc	NMCovering $(T \rightarrow S_{gt})$
boundary potential			
SE-OWT-UCM [24]	48.4	83.0	59.0
PMI-OWT-UCM [40]	47.0	86.5	61.3
semantic smoothness			
Ours w/o CRF	78.1	86.0	75.4
Ours w CRF	**80.3**	**87.3**	**76.8**

COCO-Stuff	mIoU	NMCovering $(T \rightarrow T_{gt})$	NHCovering
boundary potential			
SE-OWT-UCM [24]	30.7	43.0	32.9
PMI-OWT-UCM [40]	27.5	43.2	23.1
semantic smoothness			
Ours w/o CRF	58.7	53.5	42.1
Ours w CRF	**59.9**	**55.6**	**43.9**

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.



Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Conclusion and Future Work

Impact: This method provides a robust framework for unsupervised semantic segmentation, capable of uncovering rich, unbiased hierarchies in image data without relying on external labels or assumptions.

Applications: Suitable for use in autonomous driving, medical image analysis, and any field requiring detailed image parsing.

Future Directions:

- Adapt the method to instance and video segmentation.
- Optimize computational efficiency for real-time processing of larger input size.

NEURAL INFORMATION PROCESSING SYSTEMS

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure

Simone Rossetti Fiora Pirri

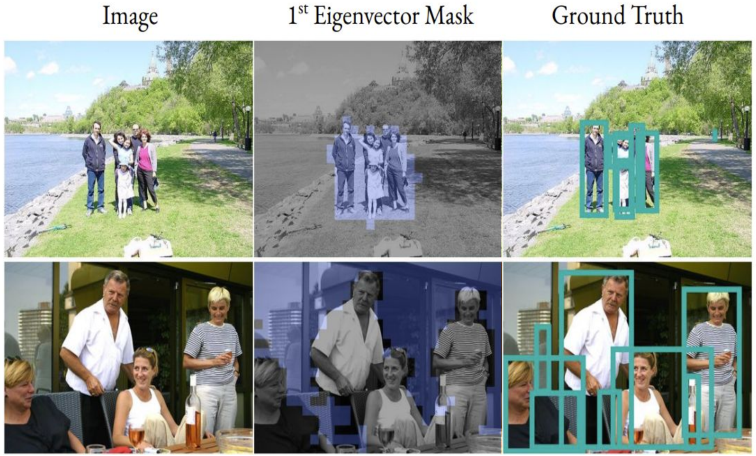
DIAG
Dipartimento di Ingegneria informatica, automatica e gestionale Antonio Ruberti

SAPIENZA
UNIVERSITÀ DI ROMA

DeepPlants

Motivation and Problem Statement (contd.)

foreground extraction



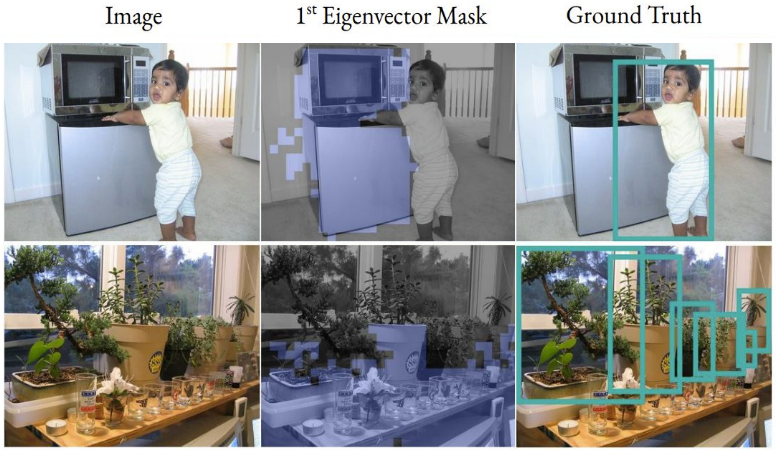


image-level single object assumption leads to inconsistencies

Deep Spectral Methods: A Surprisingly Strong Baseline for Unsupervised Semantic Segmentation and Localization, Luke Melas-Kyriazi et al., CVPR 2022.

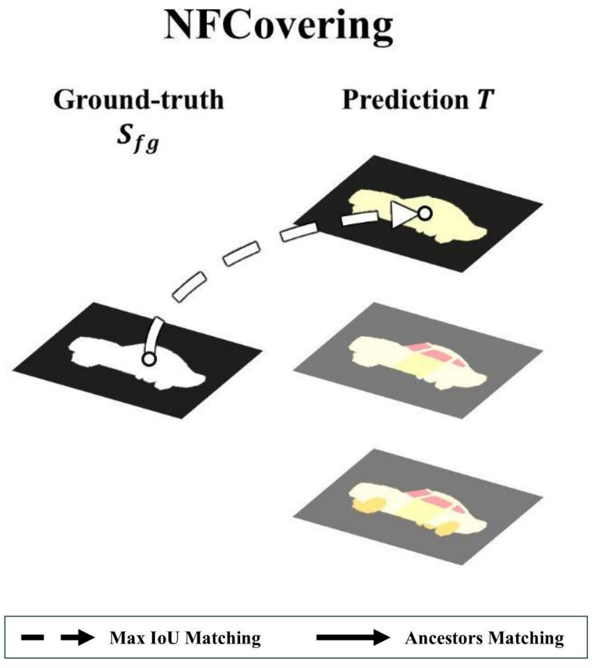
Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Evaluation of a single-level grouping

Normalised Foreground Covering (NFCovering):

- instance-level Jaccard's Index between:
  - single-level granularity of ground-truth,
  - single-level granularity of prediction.
- fails to account multiple granularities.

$$\mathrm{NFCovering}(T \to S_{fg}) := \frac{1}{|S_{fg}|} \sum_{R \in S_{fg}} \max_{R' \in T} \frac{|R \cap R'|}{|R \cup R'|}$$



Unsupervised Hierarchical Semantic Segmentation with Multiview Cosegmentation and Clustering Transformers, Ke T. et al., CVPR 2022.

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.

Technical Workflow

1. **Feature Extraction:** Uses pre-trained self-supervised models (e.g., DINO, CLIP) to derive deep feature embeddings of pixels.
2. **Iterative Graph Construction and Partitioning:**
   a. Constructs the pixels’ perturbed adjacency matrix weighted by feature similarity.
   b. Cluster pixels minimising a normalized smoothness measure on feature similarity and compute the resulting perturbation bound.
3. **Early Stopping Criteria:** Ensures that partitioning halts when granularity becomes too fine or perturbation effects become unreliable, maintaining segmentation relevance and avoiding unnecessary iterations.

Unsupervised Hierarchy-Agnostic Segmentation: Parsing Semantic Image Structure, Rossetti S. and Pirri F., NeurIPS 2024.